Between Code and Canon
Artificial Intelligence in Language and Literature Studies

Book Series

Contemporary Studies on Digital Humanities

Ali Tilbe, Yusuf Topaloğlu, Gökhan Dinar (Eds.)

Gökhan Dinar & Yusuf Topaloğlu

Between Code and Canon
Artificial Intelligence in Language and Literature Studies

maurer.press
Frankfurt am Main

Bibliographic Information of the German National Library

The German National Library lists this publication in the German National Bibliography; detailed bibliographic data are available online at the German National Library website: http://dnb.d-nb.de

© Published by Verlag Hans Jürgen Maurer

under the Creative Commons Attribution 4.0 International License (CC BY 4.0).

The work, including all of its parts, is protected by copyright. Any use beyond the narrow limits of copyright law without the consent of the copyright holder is prohibited and punishable by law. This applies in particular to reproductions, translations, and processing in electronic systems.

The book has undergone a double blind peer review.

Published in cooperation with AYG Press İstanbul and IUC University Press

maurer.press

is an imprint of
Verlag Hans-Jürgen Maurer
Im Trierischen Hof 14
60311 Frankfurt/Germany
www.maurer.press
GPSR-Kontakt: info@maurer.press

ISBN 978-3-949550-52-2
e-ISBN 978-3-949550-53-9

Table of Contents

Preface vii

Publication Ethics Statement viii

Introduction to Between Code and Canon ix

Sampling the Literary Possible: Distribution-Value and Decoding Strategies in AI-Assisted Literary Text Analysis

Alaskar Özperçin 13

The Masculine Language of Algorithms and Digital Patriarchy: Sexist Stereotypes in the Reproduction of Characters in Italian Literature

Burcu Ün & Gizem Arslan 45

Ethical Use of Gen-AI in Language Teacher Education

Nazlı Ceren Cirit-Işıklıgil 73

The Matrix of the Invisible: Distant Reading and Artificial Intelligence, and the Combinatorial Architecture of Italo Calvino's Città invisibili

Cristiano Bedin 93

Generative AI-Mediated Reception of Literary Texts in French as a Foreign Language: A Six-Stage Pedagogical Framework

Şule Çinar Yağci 119

Designing Classroom Activities and Games for Chatbot-Supported Task-Based Language Teaching

Aybegüm Aksak & Selçuk Doğan 153

A Legal Analysis of the Use of Artificial Intelligence in Türkiye: Risk, Liability and Regulatory Approach

Alkın Karakurumer 205

Beyond One-Size-Fits-All: Artificial Intelligence in Foreign Language Education for Gifted Learners

Ayşegül İşlekeller Bozca 229

From Algorithm to Voice: The Pedagogical Potential and Limitations of AI-Generated Spoken Content

Doğa Selin Geçer & Yusuf Topaloğlu..... 257

Adaptive Intelligence and the Mediating Teacher: A Critical-Pedagogical Account of Artificial Intelligence and Adaptive Learning Management Systems in Foreign Language Education

Gökhan Dinar & Ali Tilbe 301

Agentic AI in Language Education: Designing Adaptive, Personalised, and Interactive Learning Experiences

Tuncer Can & İrfan Şimşek 317

AI on the go: Technology is Reconstructing Second Language Assessment

Rabia Damla Karaçeper, Gülay Kiray & Tuncer Can 351

Comparison of Traditional Writing and ChatGPT-Assisted Writing in French as a Foreign Language (FLE)

Zeynep Büyüksaraç..... 379

Algorithmic Literature in the Digital Age: A Discussion on the Literary Competence of Artificial Intelligence

Zeyneb Dolan, Hakan Dolan 411

Preface

Few questions matter to as many people at once as the one this book addresses. Artificial intelligence now writes, reads, translates, tutors and grades, and it does so in every language classroom, every literature seminar and every examination hall that has a connection to the internet. Teachers in İstanbul and Frankfurt, in Rabat and Toronto, are making the same decisions this year about what to admit into their practice and what to keep out, and they are making them with very little help from the disciplines that ought to be guiding them. We believe that the volume in your hands speaks to that situation directly, and that it will be useful wherever language and literature are taught.

It is not, however, simply a collection of papers that happen to share a keyword. The thirteen chapters were written out of a common concern and were shaped, through invitation, review and revision, toward a common question. Their authors work on task-based teaching and agentic tutoring, on gifted learners and synthetic voices, on the computational reading of Calvino and the sexist habits of generative models, on automated assessment and on the law that will eventually govern all of this. What unites them is a refusal to treat the machine as either a miracle or a threat. Each chapter asks instead what changes for readers, learners and teachers when the instruments of interpretation are themselves made of the texts they interpret. That question, and the answers gathered here, give one of the most discussed subjects of our time a frame that we have not found elsewhere, one in which pedagogy, literary criticism, assessment and regulation are read as parts of a single problem.

The strength of the book lies in the people who wrote it. The contributors are specialists in their fields, several of them with long records of work on educational technology, Romance literatures, language testing and public law, and they brought to the project both the rigor of their disciplines and a willingness to write across them. Their chapters were subjected to editorial reading and peer review before acceptance and were revised in the light of that review. The volume is published by Maurer Press in Frankfurt am Main, in association with AYG Press İstanbul and İUC Press İstanbul, a partnership that ensures international distribution and lasting bibliographic access.

We offer the book to teachers, scholars, examiners, policymakers and graduate students alike, and we hope that it will be read, taught, contested and built upon.

Gökhan DİNAR & Yusuf TOPALOĞLU

Publication Ethics Statement

Originality and authorship: All chapters in this volume are original works that have not been published elsewhere and are not under consideration by another publisher. Where a chapter builds on a conference presentation or an earlier version, this is acknowledged in the chapter. Authorship is limited to those who have made a substantial intellectual contribution to the conception, design, analysis or writing of the chapter; all listed authors have approved the final version and accept responsibility for its content.

Editorial and review process: Each chapter was evaluated by the editors and by at least one independent reviewer with expertise in the relevant field. Reviews were conducted confidentially and impartially, and authors were asked to revise their chapters in the light of the reviewers' comments. The editors' own chapter was handled by the co-editor and an external reviewer, and the editors took no part in the evaluation of their own work.

Research ethics: Chapters reporting research with human participants confirm that the study was conducted in accordance with the principles of the Declaration of Helsinki and national regulations, that ethics committee approval was obtained where required, and that informed consent was secured from participants. Personal data have been anonymised, and the authors take responsibility for the accuracy of the data and results they report.

Use of artificial intelligence tools: Because this volume examines artificial intelligence, the editors have paid particular attention to the transparent use of generative AI in the preparation of the chapters. Generative AI tools may not be listed as authors. Where such tools were used, whether for data generation, analysis, language editing or the production of examples, the authors have disclosed the tool, its version and the purpose of its use in the relevant chapter, and they remain fully responsible for the accuracy, originality and integrity of the resulting text.

Plagiarism, citation and permissions: All chapters were screened with similarity-detection software before acceptance. Authors have cited all sources on which they rely, have obtained permission to reproduce copyrighted material, tables or figures where necessary, and have ensured that quotations are accurately attributed.

Conflicts of interest and funding: Authors have declared any financial or non-financial interests that could be perceived as influencing their work, and have acknowledged the sources of funding that supported their research, where applicable. The editors declare no conflict of interest in relation to the chapters included in this volume.

Corrections: Should a significant error or ethical concern be identified after publication, the editors and publishers will cooperate with the authors to publish a correction, an expression of concern or, where necessary, a retraction.

The Editors

Introduction to Between Code and Canon

Two words have long organised the relation of the humanities to technology. Code names the procedures through which a machine turns symbols into outputs, and canon names the body of texts, practices and judgements through which a culture decides what deserves to be read, taught and preserved. For most of the twentieth century the two words led separate lives. Large language models have brought them into the same room. A system trained on the canon now produces code-shaped language, and the language it produces returns to the classroom, the seminar and the examination hall as material that must in turn be read, taught and judged. This volume takes its title from that encounter. Its thirteen chapters ask, from different disciplinary positions, what becomes of language education and of literary studies when the instruments of reading, writing and evaluation are themselves products of the texts they are asked to interpret.

The question is not new, but its conditions are. Earlier generations of computer-assisted language learning and of computational stylistics treated the machine as an instrument whose output could be checked against a stable human standard. Generative systems unsettle that arrangement in three ways. They produce rather than retrieve, so that the boundary between a corpus and its commentary becomes porous. They adapt to the learner, so that part of the pedagogical relation is delegated to a model whose decisions remain opaque. And they carry into every output the social and ideological regularities of their training data, so that the critique of representation, long practised in literary studies, becomes a precondition of responsible use in the classroom. The contributions gathered here respond to these three displacements, and it is around them that we organise our reading of the volume, although the chapters themselves are arranged alphabetically by author so that no hierarchy of approaches is imposed on the reader.

A first group of chapters examines the machine as a reader. Cristiano Bedin turns distant reading and semantic embeddings toward Calvino's Invisible Cities to test whether computational methods can render visible combinatorial architecture that close reading intuits but rarely maps. Alaskar Özperçin moves one level deeper into the machinery and argues that decoding strategies, the settings by which a model converts probabilities into sentences, are interpretive conditions rather than technical trivia, a claim with

consequences for any study that treats generated text as evidence. Zeyneb Dolan and Hakan Dolan ask whether literary competence can be attributed to algorithmic production at all, revisiting the death of the author in the light of systems that write without a life to confess. Burcu Ün and Gizem Arslan bring the critique of representation to bear on the same systems and show how sexist stereotypes are reproduced when characters from Italian literature are regenerated by a model, reminding us that the canon a machine has absorbed is also the patriarchy it has learned.

A second and larger group addresses the machine as a participant in teaching. Aybegüm Aksak and Selçuk Doğan design classroom activities and games in which chatbots are integrated into task-based language teaching. Tuncer Can and İrfan Şimşek describe the passage from tools that answer to systems that act, and propose design principles for agentic artificial intelligence in language education. Ayşegül İşlekeller Bozca examines how artificial intelligence can serve gifted learners whose needs are routinely overlooked by uniform instruction. Doğa Selin Geçer and Yusuf Topaloğlu weigh the pedagogical potential and the limitations of machine-generated spoken content. Şule Çınar Yağcı proposes a six-stage framework for the reception of literary texts in French as a foreign language when that reception is mediated by generative tools, and Zeynep Büyüksaraç compares traditional and ChatGPT-assisted writing among learners of French. Gökhan Dinar and Ali Tilbe argue that adaptive systems deliver their most dependable gains in the rule-governed dimensions of language, while the interpretive, intercultural and ethical dimensions continue to require a mediating teacher, a position they hold to be pedagogical rather than nostalgic. Drawing on emerging AI literacy frameworks, professional guidelines, and empirical research, Nazlı Ceren Cirit-Işıklıgil proposes five discipline-specific interconnected principles—linguistic fairness, learner data protection, developmental authorship, process-visible assessment, and reflective agency—to help teacher educators equip pre-service and in-service language teachers for ethical and responsible Gen-AI use.

A third group situates these practices within the institutions that regulate them. Rabia Damla Karaçeper, Gülay Kıray and Tuncer Can trace the reconstruction of second language assessment by automated scoring, from writing to speaking, and identify the risks that travel with speed and consistency. Alkın Karakurumer offers a legal analysis of artificial intelligence

in Türkiye, examining risk, liability and the regulatory approach of the recent bill, and thereby gives the pedagogical and literary discussions a normative horizon that they too often lack.

Read together, the chapters converge on a position that we would summarise as follows. Neither enthusiasm nor refusal is an adequate response to generative systems in language and literature. What is required is a literacy that is at once technical, critical and pedagogical. It is technical because decoding parameters, training data and adaptive algorithms shape what a model can say. It is critical because the outputs inherit the inequalities of the archives they compress. It is pedagogical because the value of any tool is decided in the situated encounter between a learner, a text and a teacher who mediates between them. The volume does not claim to settle the questions it raises. It claims, more modestly, to have posed them from within the disciplines that will have to live with the answers.

This book is published by Maurer Press, Frankfurt am Main, in association with AYG Press İstanbul and İUC Press İstanbul, a partnership that we hope will carry work produced in Türkiye and its neighbouring academic communities to a wider international readership. We are grateful to the authors for their care and patience throughout the editorial process, to the reviewers whose reports improved every chapter, and to the three publishers for their trust in the project.

Sampling the Literary Possible: Distribution-Value and Decoding Strategies in AI-Assisted Literary Text Analysis

Alaskar Özperçin[1]

1. Introduction: Why Sampling Matters in Literary Analysis

Literary study has always depended on operations of selection. A critic selects a passage, isolates a motif, follows a metaphor, compares variants, or brings one interpretive frame into contact with another. Generative artificial intelligence adds a further, less visible operation of selection. When a language model responds to a literary prompt, it does not simply retrieve a complete answer waiting inside the system. It assigns probabilities to possible next tokens and uses a decoding procedure to turn those probabilities into a sequence. The resulting paragraph, interpretation, or continuation is therefore shaped not only by the source text and the prompt but also by the rules according to which some tokens remain available, some become less likely, and some are excluded from consideration.

This point gains particular significance in literary analysis. A factual question may be evaluated, at least in principle, against a determinate answer. Literary interpretation, by contrast, regularly operates in fields of underdetermination. The critic attends to ambiguity, figurative tension, unreliable perception, narrative gaps, unresolved causality, generic hesitation, and historical context. Recent studies of large language models in literary tasks likewise show that model performance varies across textual aspects and that the attribution of "understanding" becomes increasingly observer-dependent as interpretive tasks grow more complex (Jannidis et al., 2025). A

[1] Ass. Prof. Dr., Istanbul University Cerrahpaşa, Faculty of Education, Department of French Language Teaching, ORCID: 0000-0002-3674-5608, *alaskar@iuc.edu.tr*, Türkiye

study structured through Bloom's taxonomy reports substantial potential for supporting literary learning while retaining human assessment as a necessary condition for judging depth, relevance, and accuracy (Chan & Wong, 2025). The question of whether large language models can function as literary critics has accordingly entered the explicit vocabulary of literary theory (Koul, 2024). These discussions do not justify either unqualified enthusiasm or categorical dismissal. They indicate, rather, that AI-assisted literary inquiry requires methods capable of distinguishing model behavior from interpretive value.

Within this frame, decoding constitutes a hidden methodological layer. Neural generators model conditional probability distributions, but an additional procedure is required to determine how a sequence will be produced from those distributions. Decoding is external to the trained model in this practical sense, and its selection can affect the diversity and linguistic properties of generated text (Zarrieß et al., 2021). The same model and the same prompt can therefore yield markedly different outputs when the decoding strategy changes. In open-ended generation, maximum-likelihood-oriented procedures may produce blandness or repetition, whereas unconstrained sampling may admit low-probability continuations that lose coherence. Nucleus sampling was proposed in direct response to this problem, dynamically restricting generation to a high-probability region rather than sampling from the entire tail (Holtzman et al., 2020).

For literary researchers, this means that a generated interpretation cannot be treated as if it were produced by "the model" in the abstract. It is produced by a model version, a tokenizer, a prompt, a decoding configuration, a seed or platform-side random state, and a particular software environment. If these conditions are not reported, a central part of the analytical situation disappears. More importantly, the omission is not merely a reproducibility problem. It can become an interpretive error. A restrictive configuration may repeatedly select familiar causal explanations and appear to demonstrate that the model reads a fantastic tale psychologically. A more permissive configuration may produce spectral agencies, unusual metaphors, or unstable ontologies and appear to demonstrate that the model recognizes the tale's supernatural dimension. Neither inference is secure

unless the researcher asks how much of the observed tendency is attributable to the probability space allowed by decoding.

The present chapter addresses this problem by introducing distribution-value. The concept names the degree and manner in which a decoding configuration preserves, restricts, or redistributes the possibilities available at a point of generation. Distribution-value is not equivalent to diversity. A text may exhibit many distinct words while offering only one interpretive direction; another may use a narrow vocabulary while sustaining a decisive ontological hesitation. Nor is distribution-value synonymous with literary quality. A high-entropy output may be incoherent, and a highly constrained output may achieve stylistic precision. The concept is therefore organized on three interrelated but non-collapsible levels: the probabilistic field made accessible by decoding, the textual variation realized in outputs, and the interpretive possibilities constructed by those outputs.

"Véra," included in Villiers de l'Isle-Adam's *Contes cruels* of 1883, provides the worked example (Villiers de l'Isle-Adam, 1883/2020). After the death of his wife, the Count d'Athol refuses to recognize her absence and reorganizes the household around her continued presence. Objects, scents, sounds, warmth, and eventually vision itself appear to confirm what grief and desire have declared. The tale is especially productive for this exercise because it does not present ambiguity as a detachable ornament. Its fantastic-visionary structure makes uncertainty constitutive of perception, causality, and material evidence. Pollicino (2014) observes that the narrative repeatedly moves between plausible and dreamlike registers while giving the ideal and immaterial an almost tangible and audible presence. A continuation model placed at one of the tale's thresholds must therefore decide, token by token, how to negotiate absence, trace, hallucination, and return.

The generated continuations used below are not proposed as additions to Villiers's work, and they are not evaluated as autonomous literary compositions. They function as interpretive probes. Like a controlled question asked of a difficult passage, a probe makes certain tendencies observable. The generated sequence shows which lexical, causal, sensory, and ontological options become actual under stated conditions. This does not make the output an interpretation in the full scholarly sense. It makes it an object

through which the researcher can examine how probabilistic constraints interact with textual ambiguity.

The chapter proceeds from theory to application. Section 2 explains conditional generation and defines the three levels of distribution-value. Section 3 presents five decoding strategies through a reader-friendly probability example. Section 4 introduces "Véra" as a literary laboratory and selects three continuation points. Section 5 provides a reproducible workflow, including a strategy demonstration and a 27-output nucleus-sampling exercise. Section 6 connects probabilistic diagnostics to lexical diversity, semantic distance, output dispersion, and repetition. Section 7 demonstrates a researcher-led close reading. Sections 8 and 9 translate the discussion into a decision guide and a statement of limitations, ethics, and reproducibility. The concluding section returns to the chapter's central proposition: decoding parameters are not neutral controls placed outside interpretation; they help organize the space from which AI-mediated interpretation becomes possible.

2. From Probability to Literary Possibility

2.1. Language Models as Conditional Probability Systems

An autoregressive language model generates a sequence incrementally. At position *t*, it estimates a probability distribution for the next token given the preceding tokens:

$$[P(x_t x_1, , x_{t-1}).]$$

The term *token* requires emphasis. A token may correspond to a word, a part of a word, punctuation, spacing, or another unit defined by the tokenizer. The model does not initially choose among completed interpretations such as "Véra is a hallucination," "Véra is a ghost," or "the narrative suspends both explanations." It assigns numerical probabilities to possible next tokens. Those local selections accumulate into clauses, images, causal relations, and eventually recognizable interpretive orientations.

This sequential structure explains why small distributional decisions can have extensive narrative consequences. Suppose that after a description of the stopped clock sounding again, the candidate tokens include *soudain*,

pourtant, *peut-être*, *donc*, and *Véra*. Selecting *peut-être* may introduce epistemic hesitation; *donc* may encourage causal closure; *Véra* may immediately personify agency; *soudain* may intensify eventfulness without deciding its cause. Once selected, the token becomes part of the context for the next prediction. Decoding therefore does not merely choose wording for an already formed proposition. It participates in the path by which a proposition, image, or ontological commitment takes shape.

The decoding literature distinguishes probability modeling from the inference procedure that constructs the output sequence. Zarrieß et al. (2021) define decoding as the procedure that handles the search space over potential outputs and show that decoding methods may optimize likelihood, diversity, or task-specific constraints.1 This separation is methodologically valuable for literary research. It permits the researcher to ask two different questions: What distributions does the model produce in response to the passage? How does a selected decoding procedure transform those distributions into an observable text? Closed commercial systems may not expose the first question because token-level logits are unavailable. The second question nevertheless remains relevant because available parameters still condition the generated sequence.

2.2. Probability Is Not Meaning

A high-probability token is not necessarily the token that is most accurate, most literary, most faithful to the source, or most interpretively productive. Probability indicates the model's relative expectation under its training and current context. In a widely discussed paradox of neural generation, maximizing likelihood may yield dull or repetitive text even when the underlying language model performs well by standard predictive measures (Holtzman et al., 2020). The interpretive analogue is straightforward: a probable continuation can be conventional precisely because it follows dominant linguistic and narrative patterns.

This does not make low probability inherently valuable. The long tail of a model distribution contains possible surprises, but it also contains malformed words, inappropriate register shifts, factual intrusions, and semantic breakdown. Hewitt et al. (2022) describe top-*k* and top-*p* as truncation procedures that set some token probabilities to zero and interpret their function

as a form of "desmoothing," or an attempt to restrict sampling to a more plausible support.3 From a literary perspective, the difficult question is not whether restriction is good or bad. It is what kind of possibility a restriction removes, and whether that removed possibility would have constituted productive ambiguity or simple noise.

A model probability should consequently be read as a condition of textual emergence, not as a certificate of interpretation. The distinction prevents four common substitutions. Probability must not be substituted for truth; fluency must not be substituted for coherence with the source; novelty must not be substituted for interpretive force; and diversity must not be substituted for literary value. The researcher may measure each of these phenomena, but their relations remain objects of analysis rather than assumptions.

2.3. Narrative Ambiguity and the Space of Continuations

Narrative ambiguity exceeds lexical polysemy. It may arise from incomplete causality, restricted focalization, unstable perception, contradictory evidence, generic hesitation, and the postponement of closure. In "Véra," the uncertainty concerns not only what particular sentences mean but what kind of world the sentences construct. Is d'Athol's perception a symptom of grief? Does his sustained desire animate the household as a shared illusion? Do objects acquire evidentiary force? Does Véra return? Or does the narrative temporarily render these alternatives inseparable?

The tale repeatedly distributes signs across sensory channels. A voice is heard, a scent is perceived, warmth remains in pearls, blood appears fresh, flowers revive, the lamp relights, the stopped clock sounds, and the key of the tomb finally falls into the room. Pollicino (2014) reads this structure through threshold relations—love and death, interior and exterior, light and darkness—and emphasizes the gradual amplification of the sensory sphere.12 Such a narrative is particularly suitable for continuation analysis because each threshold can be extended in several directions. A model may psychologize the sign, supernaturalize it, materialize it without explanation, convert it into metaphor, or dissolve it into stylistic incoherence.

Here the space of possible continuations should not be imagined as a list of complete endings. It is better understood as a branching field in which local choices redistribute later possibilities. An epistemic marker such as *comme*

si can suspend commitment; an indicative declaration can close it. A sensory verb can keep the event within focalized experience; an omniscient causal statement can move outside that experience. A concrete noun can provide an object around which presence gathers; an abstract noun can redirect the passage into commentary. The literary significance of decoding emerges because the algorithm regulates access to these local moves.

2.4. Defining Distribution-Value

> Distribution-value refers to the extent to which a decoding configuration preserves, restricts, or redistributes the lexical, semantic, and interpretive possibilities available at a given point of literary continuation.

The expression includes *value* not to imply a single score but to insist that distributional breadth becomes meaningful only in relation to a research purpose. A configuration that is valuable for source-faithful paraphrase may be unproductive for mapping competing ontologies. A permissive configuration useful for discovering unusual associations may be unsuitable for a reproducible classroom demonstration. Value is relational, and its relation must be stated.

Level of distribution-value	Primary object	Possible observations	Interpretive caution
Probabilistic	The token field retained or reshaped by decoding	Entropy, finite candidate set, nucleus size, selected-token surprisal	Logit access may be unavailable; token-level breadth is not textual quality.
Textual	The lexical and semantic variation realized in outputs	MTLD, HD-D, Distinct-1/2, repetition, cosine distance, pairwise dispersion	Metrics are length- and model-sensitive and may reward incoherence.
Interpretive	The way an output organizes literary uncertainty	Ontological orientation, causal language, epistemic markers, sensory evidence, closure	Requires situated close reading; cannot be reduced to a composite score.

The levels interact but should not be collapsed. A large nucleus may produce a small range of realized texts by chance or because the model distribution remains sharply concentrated. High lexical diversity may coexist

with a single causal explanation. High semantic distance from the source may indicate interpretive expansion, generic drift, or total breakdown. A composite distribution-value index would conceal these differences by placing incommensurable observations on a single scale.

Distribution-value is therefore best treated as a profile. The probabilistic profile shows what the decoder made available at token level. The textual profile shows what was realized across generated samples. The interpretive profile shows what those realizations do to the literary problem. In this three-level arrangement, measurement prepares the object for reading, while reading returns the measures to the source text and research question.

3. A Reader-Friendly Guide to Decoding Strategies

Decoding strategies can appear remote from literary practice when explained only through implementation details. The following comparison uses a shared imaginary distribution. After a passage describing a faint movement in Véra's empty room, suppose that the model assigns the next-token probabilities shown below.

Candidate token	Initial probability	Possible literary function
une	.28	Continues a nominal description without deciding agency.
Véra	.22	Introduces direct agency and may strengthen supernatural presence.
peut-être	.16	Marks uncertainty and preserves epistemic hesitation.
soudain	.13	Intensifies eventfulness while leaving cause open.
pourtant	.09	Establishes contrast and may resist the preceding interpretation.
ombre	.06	Moves toward figurative or spectral imagery.
preuve	.04	Risks premature evidentiary closure.
ailleurs	.02	Opens a spatial displacement with weak local support.

This distribution is pedagogical, not extracted from the model used later. Its purpose is to make each algorithm's operation visible on the same field.

3.1. Greedy Decoding as a Baseline

Greedy decoding selects the highest-probability token at every step. In the table, it selects *une*. At the next step, it again selects the most probable token under the updated context, and the process continues until a stopping condition is reached. Greedy decoding is deterministic when the model and environment are fixed. It is therefore useful as a baseline: it shows one locally maximal path through the distribution.

The strength of greedy decoding is also its limitation. Local probability maxima do not guarantee the most coherent complete sequence, and repeated preference for high-probability continuations may produce formulaic phrasing. Holtzman et al. (2020) connect likelihood-oriented decoding with blandness and repetition in open-ended generation. In literary work, a greedy continuation can reveal the model's most conventional route through a passage. It should not be mistaken for the model's single interpretation. It is one path produced by one rule of selection.

For a critic, the baseline question becomes: What happens to ambiguity when every step favors the locally safest continuation? If epistemic markers consistently lose to declarative constructions, the output may close uncertainty quickly. If familiar genre terms dominate, the continuation may convert a complex fantastic structure into a recognizable ghost story or grief narrative. Such tendencies can be informative, provided that they are attributed to the interaction of model and decoder rather than to the model alone.

3.2. Temperature

Temperature rescales logits before probabilities are calculated:

$$[P_{}(x_i)=.]$$

When (<1), the distribution becomes sharper: high-probability tokens gain relative weight and low-probability tokens lose it. When (>1), the distribution becomes flatter: lower-probability tokens become more competitive. Temperature does not, by itself, remove tokens from the vocabulary. It redistributes their relative probabilities.

This distinction matters because temperature is often treated as a direct "creativity" control. High temperature may increase variability, but creativity is not a mathematical consequence of flattening a distribution. The procedure may admit a productive metaphor, an anachronistic concept, a malformed word, or a register-breaking phrase. Likewise, low temperature may produce precision or cliché. Temperature changes the conditions of selection; it does not evaluate the literary consequence.

In the imaginary distribution, lowering temperature would strengthen *une* and *Véra* relative to *ombre* and *ailleurs*. Raising it would reduce the dominance of the first two and give the tail a larger share. For a study of narrative ambiguity, temperature may therefore be used to test how stable an interpretive orientation remains when the probability contrast among candidates is sharpened or relaxed. The researcher should vary temperature separately before combining it with truncation, because changing temperature and top-*p* simultaneously makes attribution difficult.

3.3. Top-k Sampling

Top-*k* sampling retains only the *k* highest-probability tokens, sets the rest to zero, renormalizes the retained probabilities, and samples from them. If (k=4) in the example, *une*, *Véra*, *peut-être*, and *soudain* remain available; *pourtant*, *ombre*, *preuve*, and *ailleurs* are excluded.

Top-*k* creates a fixed-size candidate pool. This property makes the method intuitive, but a fixed number does not adapt to the shape of the distribution. At one step, probability may be concentrated in three plausible tokens; at another, twenty tokens may be reasonably supported. A fixed (k) can be too permissive in the first context and too restrictive in the second. Hewitt et al. (2022) place top-*k* among truncation strategies that attempt to recover a plausible support by removing low-probability mass.

For literary analysis, top-*k* can be useful when the researcher wants a transparent and constant cap on local alternatives. It can make demonstrations easy to explain: *k* tokens enter the pool, and all others leave it. Yet comparisons across passages require care. The same (k) does not imply the same retained probability mass, uncertainty, or semantic breadth. top_k=40 at a highly predictable punctuation point is not distributionally equivalent to top_k=40 at a narratively open threshold.

3.4. Nucleus Sampling

Nucleus, or top-*p*, sampling retains the smallest set of highest-probability tokens whose cumulative probability reaches a threshold (p):

$$[\ V\_p=\{V'V:\_\{xV'\}P(x)p\}.\ ]$$

At (p=.80) in the example, the cumulative mass reaches .79 after the first four tokens and .88 after *pourtant*. The nucleus therefore contains five tokens. At another generation step, two tokens might be sufficient to reach .80; elsewhere, fifty might be needed. This dynamic adjustment distinguishes nucleus sampling from top-*k*. Holtzman et al. (2020) propose the method as a way to sample from a shifting high-probability region while avoiding the unreliable tail.

Nucleus sampling is especially suitable for the present walkthrough because its threshold can be interpreted as a controlled relaxation of the candidate field. A restrictive level such as .50 often yields a small nucleus and favors conventional local choices. A moderate level such as .80 admits more alternatives while retaining substantial truncation. A permissive level such as .95 may include a much larger set, particularly in uncertain contexts.

The threshold should not be translated directly into a literary category. top_p=.95 does not mean that 95% of literary meaning has been preserved, and top_p=.50 does not mean that half the vocabulary remains. The parameter refers to cumulative model probability at each token step. Moreover, Hewitt et al. (2022) show that top-*p* may behave counterintuitively in some sharply peaked distributions, even excluding plausible alternatives when one token dominates. Its value lies in adaptive truncation, not in a guarantee of optimal support.

3.5. Locally Typical Sampling

Locally typical sampling starts from an information-theoretic question. A token's surprisal is the negative logarithm of its probability. The distribution's conditional entropy represents expected information content. Meister et al. (2023) define locally typical sampling by favoring tokens whose information content is close to that expectation, then sampling within a set determined by a probability-mass parameter.4 Their evaluations in

summarization and story generation report competitive quality and reduced degenerate repetition relative to nucleus and top-*k* sampling.

The intuitive contrast is important. Nucleus sampling orders tokens primarily by probability and includes enough of the most probable tokens to meet a cumulative threshold. Locally typical sampling orders them by how close their surprisal is to the distribution's expected surprisal. A very probable token can be less "typical" if it is substantially less informative than the local expectation; a somewhat less probable token can be more typical if its information content lies near that expectation.

For literary work, this procedure suggests a way to avoid equating probability rank with appropriateness. It may preserve locally informative choices without opening the entire tail. Yet the same caution applies: statistical typicality is not stylistic fidelity, historical appropriateness, or interpretive subtlety. It is an information-theoretic relation within the model distribution. The literary researcher should treat it as another way of shaping the probe, not as a theory of literariness.

3.6. Parameter Interaction

Generation interfaces often expose temperature, top-*k*, and top-*p* together. The order of operations generally involves rescaling, filtering, renormalizing, and sampling, although exact implementation details should be checked for the software version in use. When temperature and a truncation threshold change together, the observed output cannot be attributed cleanly to either parameter. Flattening the distribution can enlarge the number of tokens needed to reach a nucleus threshold; sharpening it can reduce that number. Thus, temperature=1.2, top_p=.95 is not simply a "more creative" form of temperature=.8, top_p=.80. It is a joint transformation of probability ratios and support.

A practical demonstration should therefore begin by changing one factor at a time. The full walkthrough in Section 5 fixes temperature at .80 and varies only top-*p*. Seeds are repeated across levels. This does not eliminate model-specific behavior, but it makes the relation between threshold and observed output more inspectable.

3.7. Common Misconceptions

> Methodological warning. High temperature is not identical with creativity; high top-p is not identical with literary quality; low top-p does not guarantee coherence; top_k=50 is not equivalent to top_p=.90; and results from different models cannot be interpreted as parameter effects alone.

These misconceptions share a common error: they convert a distributional operation into an aesthetic judgment. Decoding parameters describe how a model moves through a probability field. Literary evaluation begins only after that operation has produced text and the text has been returned to a historically and formally situated act of reading.

4. "Véra" as a Literary Laboratory

4.1. The Narrative Situation

"Véra" begins after burial. The Count d'Athol returns to the room in which his wife died, surveys the objects she has left, and finds ordinary futurity impossible. He recalls a union so intense that spirit and body seemed to interpenetrate. From this point, grief becomes domestic practice. D'Athol orders the servant Raymond to continue serving "the countess and me," dismisses the other servants, and establishes a closed household in which Véra's death is not acknowledged. Raymond initially interprets the performance as a protective accommodation to madness, but the repeated routine gradually affects him as well.

The narrative then moves from willed illusion to ambiguous manifestation. A dress is glimpsed, a voice is heard, a bell rings, a scent is perceived, and the atmosphere appears to acquire the form of presence. On the first anniversary of Véra's death, material signs intensify. Pearls retain bodily warmth, blood appears fresh, the icon lamp relights, flowers bloom, and the stopped clock sounds. Véra becomes visible and speaks. Yet the vision collapses when d'Athol remembers aloud that she is dead. At the end, the silver key he had thrown inside the sealed tomb falls from the bed. The object seems to answer his request for a path back to her, but its evidentiary status remains inseparable from the narrative machinery of desire, faith, and fantastic hesitation (Villiers de l'Isle-Adam, 1883/2020).

4.2. Why "Véra" Is Suitable

The tale is productive for AI-assisted continuation because it places psychological and supernatural explanations in sustained proximity. D'Athol's conduct can be described through grief, denial, autosuggestion, ritual, and altered perception. At the same time, the story progressively externalizes signs that exceed his private consciousness. Raymond perceives some of them; objects seem to participate; the final key introduces a material trace that resists easy reduction.

Pollicino (2014) characterizes "Véra" as a fantastic-visionary narrative and emphasizes its movement across liminal oppositions. The story does not simply alternate between an objective world and a subjective mistake. It constructs a chamber in which interior conviction appears to saturate exterior matter. This makes continuation points methodologically valuable. A model can preserve focalized uncertainty, select a causal explanation, displace the scene into another genre, or repeat the source. Each result reveals something about the interaction among passage, prompt, model, and decoder.

4.3. A Semiotic Map of Presence and Absence

The following four-part map guides close reading. It is not a coding scheme and does not produce frequency counts.

Position	Narrative realization	Question for generated continuations
Presence	Véra appears as bodily, speaking, and externally perceptible.	Does the output assert that she has returned as an agent?
Absence	Death, burial, silence, emptiness, and the widowed room.	Does the output restore loss as final and factual?
Non-presence	Memory, hallucination, dream, role-play, or autosuggestion.	Does it explain signs as products of mind or ritual?
Non-absence	Trace, scent, warmth, voice, moved object, or spectral pressure.	Does it preserve signs that resist both full presence and simple absence?

The concept of **non-absence** is especially useful. The warm pearl is not Véra's body; neither is it mere emptiness. It is a trace whose status remains

open. A generated continuation that turns every trace immediately into a ghost reduces the semiotic interval. One that dismisses every sign as hallucination performs a different closure. The most interesting outputs may not be the most ornate but those that sustain movement among these positions.

4.4. Selecting Continuation Points

Three passages were selected from the Project Gutenberg text of *Contes cruels* (Villiers de l'Isle-Adam, 1883/2020). **P1** ends with d'Athol's rhetorical refusal of death: "Is the soul of the cellos carried away in the cry of a string that breaks?" **P2** ends when imaginary and real become indistinguishable and a form attempts to weave itself into indeterminate space. **P3** ends when the clock stopped for a year strikes again after warmth, blood, light, and flowers have animated the room.

Passage	Interpretive pressure	Reason for selection
P1: Denial of death	Metaphysical analogy competes with bereavement.	Tests whether the continuation psychologizes denial, extends figurative reasoning, or asserts survival.
P2: Intensification of perception	Presence emerges without stable embodiment.	Tests epistemic markers, sensory modalities, and the treatment of non-absence.
P3: Animation of objects	Material signs approach evidence.	Tests causal closure, agency, and the transition from trace to manifestation.

The full passages are supplied in the generation metadata and supporting script. Their boundaries are fixed before generation; they are not selected retrospectively to favor a particular result.

5. The Hands-On Workflow

5.1. Choosing and Reporting a Model

A decoding study should use a model that exposes the parameters under investigation. Closed chat interfaces may allow temperature or top-*p* adjustments while withholding top-*k*, locally typical sampling, logits, tokenizer details, or seeds. Such systems remain usable for some forms of

inquiry, but their limits should be reported rather than silently replaced by assumptions.

The worked example uses Qwen/Qwen2.5-0.5B-Instruct, a 0.49-billion-parameter, instruction-tuned model with multilingual support that was executed locally through Transformers (Qwen Team, 2024).14 The point is not to recommend this model for literary scholarship. Its small size makes its weaknesses conspicuous and keeps the demonstration computationally reproducible. All 32 records—five Module A outputs and twenty-seven Module B outputs—were generated on August 27, 2026, with Python 3.12.3, PyTorch 2.13.0, Transformers 5.16.1, and a maximum of 96 new tokens. The raw CSV, metrics, prompt, source passages, seeds, and scripts accompany this chapter.

This choice demonstrates an important methodological principle. Model adequacy and parameter transparency can conflict. A larger proprietary model may produce more fluent French but reveal less about its distribution. A small open model may expose logits and sampling support while producing stylistic failures. Distribution-value analysis should not hide this tradeoff. The model is part of the finding situation, not a neutral instrument outside it.

5.2. Preparing a Neutral Prompt

The prompt asks for a French continuation consistent with the narrative voice and atmosphere. It prohibits summary and commentary, specifies an approximate length, and asks the model to preserve rather than resolve uncertainty between psychological and supernatural explanations. The final instruction is a deliberate control. Without it, a chat-tuned model may treat the task as an invitation to explain the passage. With it, the prompt still does not prescribe whether a sign is psychological or supernatural.

> Continue the following passage in French in a manner consistent with its narrative voice and atmosphere. Do not summarize the passage or explain your choices. Preserve rather than resolve its uncertainty between psychological and supernatural explanations. Produce approximately 80–110 French words.

No prompt is perfectly neutral. “Consistent,” “atmosphere,” and “preserve uncertainty” are interpretive instructions. Neutrality here means that the prompt does not select one of the competing ontological explanations in

advance. A different research question could justifiably use a different prompt, but prompt versioning would then be required.

5.3. Module A: Strategy Demonstration

Module A applies five strategies to P2 with the same seed wherever randomness is involved. The settings are pedagogical contrasts rather than optimized configurations: greedy decoding; unrestricted high-temperature sampling at 1.20; top-*k* sampling with k=40 and temperature .80; nucleus sampling with top_p=.80 and temperature .80; and locally typical sampling with typical_p=.80 and temperature .80.

The resulting texts should be read as diagnostic artifacts. Greedy decoding largely repeats the source. High-temperature unrestricted sampling produces multilingual fragments and invented forms, demonstrating that a flat, untruncated distribution does not constitute creativity. Top-*k* produces a more grammatical but abstract commentary on "mécanismes psychologiques et matériels." Nucleus sampling shifts into metadiscursive exposition about characters and communication. Locally typical sampling remains more coherent but loses the source's style and moves toward a generic account of solitude.

Strategy	Mean token entropy	Mean finite support	Observed behavior in this run
Greedy	0.20	151,936	Source copying and incomplete continuation.
Temperature 1.20, no truncation	7.13	151,936	Severe lexical and multilingual degeneration.
Top-*k* = 40	1.75	40.64	More controlled syntax, but analytical paraphrase replaces narration.
Top-*p* = .80	1.41	17.31	Moderate local support; coherence improves but metadiscourse appears.
Typical-*p* = .80	1.39	23.94	Comparable informational restriction; generic narrative continuation.

The fractional support values are means across token steps. Special-token and implementation effects can produce stepwise differences around the nominal filter size. More importantly, these diagnostics describe the processed sampling distribution in this implementation. They should not be compared uncritically with logits from another library or platform.

Module A performs a useful negative demonstration. If only the most fluent output had been retained, the researcher would conceal evidence about copying, instruction failure, and distributional degeneration. Keeping the failures protects the distinction between a sampling mechanism and its hoped-for literary effect.

5.4. Module B: Complete Nucleus-Sampling Walkthrough

Module B narrows the question. Temperature remains fixed at .80, top-*k* is disabled, and only top-*p* varies. Three passages, three top-*p* levels, and three seeds produce twenty-seven outputs:

$$[3\ p\ =27\ .]$$

The levels are .50 (restrictive), .80 (moderate), and .95 (permissive). Seeds 11, 29, and 47 recur at every level. Reusing seeds does not ensure semantically aligned decisions once distributions change, but it documents the random initialization and supports rerunning the procedure.

The design does not support inferential claims. Nine texts per level are enough to demonstrate the workflow, not enough to establish universal relations between top-*p* and literary interpretation. The passages are not independent samples from a literary population, the model is singular, and the outputs share prompts and source material. Descriptive aggregation is used only to make patterns available for inspection.

5.5. Recording Outputs

Each output receives an identifier and is stored without correction. The record includes module, passage, strategy, temperature, top-*p*, top-*k*, typical-*p*, seed, model, prompt version, UTC timestamp, token limit, generated-token count, word count, token diagnostics, and raw generated text. A minimal schema appears below.

```
output_id, module, passage_id, strategy, temperature,
top_p, top_k, typical_p, seed, model, prompt_version,
```

```
generated_at_utc, max_new_tokens, generated_token_count,
word_count, generated_text
```

Unedited storage is essential because correction changes the object of analysis. Removing a repeated clause lowers measured repetition; fixing agreement improves perceived style; deleting metadiscourse conceals instruction failure. If a cleaned reading version is needed for publication, it should be stored in a separate field and never replace the raw output.

5.6. Reproducibility Checklist

Question	Required record
Was the same model used?	Repository name, revision or checksum when available, and model configuration.
Was the prompt stable?	Full system and user prompts with version identifiers.
Was randomness recorded?	Seed values and framework-level deterministic settings.
Were outputs altered?	Raw output archive plus a transparent correction policy.
Were lengths controlled?	Token limits, stopping rules, realized token counts, and word counts.
Can the run be dated?	Generation timestamp and software environment.

Reproducibility is not identity across all hardware. Floating-point behavior, library changes, model-repository revisions, and chat-template updates may affect output. The purpose of complete reporting is to make differences diagnosable rather than to promise an impossible permanence.

6. Measuring Distribution-Value

6.1. Lexical Diversity

Lexical diversity measures ask how varied the realized vocabulary is, but different indices operationalize variation differently. Type-token ratio is strongly affected by text length. MTLD estimates the mean length of token sequences that maintain a specified level of lexical variation, while HD-D derives diversity from the hypergeometric distribution. McCarthy and

Jarvis (2010) validate MTLD and HD-D and recommend considering multiple indices rather than relying on a single measure.

For short generated continuations, MTLD is useful as a comparative prompt but remains unstable when outputs are incomplete, copied, or malformed. HD-D requires an adequate token sample and should be marked missing rather than forced on extremely short texts. Language-specific tokenization also matters. Apostrophes, elisions, hyphenation, and invented forms can inflate or distort counts in French.

```
from lexical_diversity import lex_div as ld
tokens = tokenize_french(generated_text)
mtld = ld.mtld(tokens)
hdd = ld.hdd(tokens) if len(tokens) >= 42 else None
```

6.2. Semantic Distance from the Source

A sentence-embedding model maps the source passage and continuation into vectors. Their cosine distance is:

$$[d_{}(x,y)=1-.]$$

Sentence-BERT demonstrates how semantically meaningful embeddings can be compared with cosine similarity (Reimers & Gurevych, 2019). The walkthrough uses a multilingual Sentence-Transformers model because the passages and outputs are French. A larger distance indicates that the continuation occupies a more distant region in that embedding space. It does not reveal whether the movement is a productive extension, a stylistic transformation, or nonsense.

6.3. Dispersion Among Continuations

Source distance and output dispersion answer different questions. A set of outputs may all remain close to the passage yet differ from one another in small but significant causal choices. Alternatively, all may drift in the same direction and show high source distance with low internal dispersion. Pairwise cosine distance among continuations generated under the same passage and parameter level makes this distinction visible.

The present walkthrough computes, for each continuation, its mean distance from the other outputs in the same passage-by-top-*p* group. These values are

then described across levels. Because each group contains only three seeds, the measure is illustrative and highly sensitive to a single failure.

6.4. Repetition and Formulaicity

Distinct-1 and Distinct-2 calculate the proportion of unique unigrams and bigrams. Distinct-*n* entered generation evaluation as a way to quantify lexical diversity in responses (Li et al., 2016). Later work shows that the original metric is biased by sequence length and proposes expectation-adjusted alternatives (Liu et al., 2022). A responsible workflow therefore reports realized lengths and treats Distinct-*n* as one observation among several.

Repetition can also be measured through the proportion of bigram positions belonging to repeated bigram types. This captures overt recurrence but misses conceptual formulaicity. An output that says "world," "reality," "existence," and "phenomenon" only once each may still be generically abstract. Conversely, intentional literary repetition may be formally meaningful. Automatic repetition measures cannot decide between stylistic pattern and degeneration.

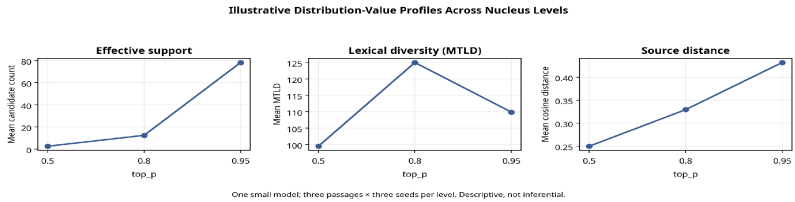
6.5. Token-Level Measures

When processed scores are available, three diagnostics are especially instructive. **Token entropy** estimates uncertainty in the available distribution. **Effective support size** counts finite candidates after filtering in the present implementation. **Selected-token surprisal** measures the information content of the realized token under that processed distribution. Locally typical sampling is directly motivated by the relation between token information content and conditional entropy (Meister et al., 2023).

These values depend on where they are measured. Entropy before temperature scaling differs from entropy after scaling; support before top-*p* differs from support after truncation. The script must state whether its scores are raw logits, processed logits, or renormalized probabilities. In the worked script, diagnostics are calculated from the score tensors returned after the configured generation processors.

6.6. Descriptive Profile from the Worked Example

top_p	Out-puts	Mean effective support	Mean MTLD	Mean source distance	Mean within-group dispersion
.50	9	2.61	99.49	.250	.324
.80	9	12.47	124.99	.329	.491
.95	9	78.26	109.89	.432	.505



Illustrative Distribution-Value Profiles Across Nucleus Levels

The profile makes three observations available. First, higher top-*p* substantially enlarges the processed candidate field in this run. Second, semantic distance and dispersion rise across the three levels. Third, lexical diversity does not increase monotonically: it peaks at .80 and falls at .95. This last pattern is methodologically useful. A broader support does not automatically realize a richer text. At .95, several outputs contain malformed or contextually displaced phrases, which may reduce coherent lexical development even as semantic distance increases.

No statistical test is applied. The values are dependent on this model, these passages, these seeds, this tokenizer, and these metric implementations. They prepare questions for close reading: What kind of distance appears at .95? Does dispersion represent alternative readings or unrelated drift? Why does the restrictive condition remain close to the source—because it preserves style, or because it copies?

6.7. Why the Measures Should Not Be Collapsed

A composite distribution-value score would reward some mixture of entropy, diversity, distance, and low repetition. Such a score would conceal incompatible outcomes. High distance can signal imaginative extension or

loss of the passage. High MTLD can signal lexical richness or a sequence of unrelated terms. Low repetition can indicate control or simply brevity. Large support can preserve possibility or expose noise.

The measurements should instead remain aligned in a profile. Their function is diagnostic: they indicate where reading effort should be directed. Literary value enters only through an account of what the continuation does to the source's formal and interpretive problem.

7. Guided Close Reading of Generated Continuations

7.1. A Close-Reading Grid

Researcher-led close reading replaces independent coding in this practical chapter. The purpose is not to estimate category frequencies but to demonstrate how an output may be returned to the literary passage and the decoding condition that produced it.

Dimension	Guiding question
Ontological orientation	Is Véra constructed as bodily, spectral, mental, absent, or unresolved?
Causality	Are events explained psychologically, materially, religiously, or supernaturally?
Epistemic language	Do markers such as *sembler*, *peut-être*, *comme si*, or conditional forms suspend commitment?
Perception	Which sensory channels create an effect of reality?
Narrative closure	Does the continuation resolve, defer, or transform ambiguity?
Formal relation	Does it continue narration, repeat the source, shift into commentary, or change genre?

The final dimension is essential for chat-tuned models. A response may discuss ambiguity rather than narrate it. Such metadiscourse can contain relevant concepts while failing the continuation task. Instruction adherence is therefore part of the literary-methodological reading.

7.2. Restrictive Nucleus Example

A P2 output at top_p=.50 and seed 47 begins:

> Il y avait quelque chose qui semblait se dérouler sans cause, comme si la vie d'Athol était interrompue par quelque chose de mystérieux. La jeune femme … n'avait pas vraiment disparu.

The sequence contains two useful epistemic devices: *semblait* and *comme si*. They place the event inside a field of appearance rather than direct assertion. "Without cause" also suspends explanation. Yet the continuation quickly declares that the young woman "had not really disappeared," moving from hesitation toward non-absence. It does not fully materialize Véra as bodily presence, nor does it reduce her to memory. In this respect, the restrictive condition maintains an intermediate ontology.

Its weakness appears in repetition and incompletion. The indefinite *quelque chose* recurs, and the output later returns to activities already present in the passage. The relatively small mean support at the restrictive level helps explain—not prove—this conservative movement. Close reading distinguishes two different effects hidden by low source distance: preservation of ambiguity through epistemic markers and failure to advance the narrative through repetition.

7.3. Moderate Nucleus Example

A P3 output at top_p=.80 and seed 11 continues the animation of objects:

> Le comte, sans bouger du collier, tourna lentement la page en silence, s'attendant à entendre la musique finir mais n'y croyant pas. Il resta silencieux pendant plusieurs minutes, absorbé par la symphonie qui enveloppait son esprit.

This output neither explains the sounding clock nor asserts Véra's return. It transfers agency from the unexplained room to the count's attention. The expectation "to hear the music end" is qualified by disbelief, producing an epistemic tension between sensory anticipation and rational resistance. The "symphony" that envelops his mind shifts the sign toward non-presence: a mental and aesthetic experience rather than an external apparition.

Formally, the continuation is more successful than many outputs in the run. It continues in narrative prose, maintains the scene's acoustic field, and avoids overt commentary. At the same time, "sans bouger du collier" is

awkward, and the clock's impossible action is displaced rather than developed. The moderate nucleus supports a coherent deferral, but it does not independently demonstrate that .80 is optimal. It provides one instance in which expanded support and retained constraint happen to align with the passage's hesitation.

7.4. Permissive Nucleus Example

A P2 output at top_p=.95 and seed 11 begins:

> Dans son domaine céleste, Athol rêvait d'une existence comme celle de sa maîtresse, sans égard pour la réalité terrestre. Son imagination … découvrant des écosystèmes et des civilisations antérieures à sa propre.

The first phrase intensifies metaphysical separation and could, in another continuation, develop the contrast between earth and transcendence. The later reference to "ecosystems" and prior civilizations, however, has no prepared relation to the chamber, Véra's sensory trace, or Villiers's narrative lexicon. Semantic distance is high because the output has not merely opened an alternative interpretation; it has changed the conceptual field.

This distinction exemplifies why permissiveness cannot be equated with interpretive expansion. The output enlarges ontology but weakens textual motivation. Its novelty is topical rather than hermeneutic. The continuation does not negotiate presence and absence; it leaves their scene. A metric registers distance, whereas close reading identifies narrative disintegration.

7.5. Linking Metrics and Interpretation

Across the worked sample, mean semantic distance rises from .250 at top-p .50 to .432 at .95. Mean within-group dispersion likewise rises. These observations justify inspecting permissive outputs for alternative developments. They do not show that the developments are valuable. The restrictive example uses epistemic markers but repeats the source; the moderate example sustains acoustic uncertainty; the permissive example introduces concepts unrelated to the passage.

> A higher semantic distance indicates departure from the source passage, but only close reading can determine whether that departure constitutes interpretive expansion or narrative disintegration.

Metrics and close reading therefore enter a reciprocal relation. Metrics locate variation; close reading differentiates its functions. Close reading may then expose a measurement problem—for example, copying that falsely appears source-faithful or malformed novelty that inflates diversity—leading the researcher to revise the metric profile. The procedure is iterative rather than hierarchical.

8. A Practical Decision Guide for Literary Researchers

No decoding setting is universally appropriate. The initial configuration should follow the analytical purpose, and the researcher should retain at least one comparison condition capable of revealing what the preferred setting excludes.

Research purpose	Reasonable starting configuration	Main risk	Necessary comparison
Source-near explanation	Low temperature; low-to-moderate top-*p*	Formulaicity, copying, premature closure	Greedy baseline and one more permissive sample.
Alternative interpretive hypotheses	Moderate temperature; moderate-to-high top-*p*; multiple seeds	Drift mistaken for insight	Source-distance profile plus close reading.
Formal imitation	Controlled temperature and nucleus sampling	Surface resemblance without narrative logic	Compare syntax, focalization, and imagery, not vocabulary alone.
Mapping ambiguity	Several top-*p* levels with fixed model, prompt, and temperature	Parameter effect mistaken for textual meaning	Repeated seeds and explicit ontological grid.
Surreal or experimental continuation	Locally typical or permissive nucleus sampling	Genre change and narrative disintegration	Restrictive control and instruction-adherence check.
Reproducible demonstration	Open-weight model; fixed revision, environment, prompt, and seeds	Version and hardware variation	Archive raw outputs and metadata.

A practical sequence follows from this table. First, define the literary problem without reference to a model—for example, how a passage sustains uncertainty between trace and apparition. Second, choose a continuation point where that problem is active. Third, select one parameter to vary while holding the others constant. Fourth, generate multiple outputs and

preserve failures. Fifth, examine probabilistic and textual profiles. Finally, conduct close reading with explicit attention to causality, epistemic language, perception, and closure.

The strategy should be revised when the outputs fail to answer the research question. If the model copies, increase the length of the instruction boundary, add a non-repetition constraint, or select a stronger model before interpreting parameter differences. If all outputs explain rather than narrate, revise the prompt. If permissive levels produce multilingual fragments, the appropriate conclusion concerns the model-decoder interaction, not the literary text.

9. Limitations, Ethics, and Reproducibility

9.1. Model, Prompt, and Decoder Effects

The principal limitation is confounding. An output results from the model, its training distribution, tokenizer, instruction tuning, prompt, context length, chat template, decoding implementation, and random state. Parameter effects cannot be generalized from one model. Even within one model family, a revision may alter logits or safety behavior. The worked example therefore supports a method, not a ranking of top-p levels.

Prompt effects are equally consequential. The instruction to preserve uncertainty encourages epistemic restraint. A prompt asking for "the most plausible explanation" would reorganize the task before decoding begins. Researchers should treat prompt and decoder as distinct but interacting controls and should report both.

9.2. Closed Interfaces and Hidden Distributions

Commercial interfaces may update models without notice, conceal token scores, cap parameter ranges, or apply undocumented post-processing. In such cases, the researcher may still compare observable outputs, but claims about entropy, support size, or selected-token surprisal are unavailable. The absence of logits should be stated explicitly. Estimated distributions reconstructed from text are not equivalent to model probabilities.

9.3. Diversity Is Not Literary Quality

The walkthrough shows that wider support can increase semantic distance while reducing coherent lexical development. A malformed continuation may score as diverse; a precise continuation may remain close to the source. Literary quality is historically, formally, and interpretively situated. It cannot be inferred from high entropy, novelty, or distance.

This caution extends to AI "understanding." Jannidis et al. (2025) argue that increasingly complex literary tasks make evaluation dependent on how human observers interpret model behavior. Chan and Wong (2025) likewise retain expert human judgment when assessing depth, accuracy, and relevance. Generated continuations should therefore assist inquiry without being substituted for the critic's account of evidence.

9.4. Attribution, Copyright, and Scholarly Integrity

AI continuations must not be presented as writing by Villiers de l'Isle-Adam. They should be labeled with model, date, and generation conditions. Public-domain source texts are preferable for reproducible demonstrations because passages and prompts can be shared. Project Gutenberg identifies *Contes cruels* as public domain in the United States and supplies the French text used here (Villiers de l'Isle-Adam, 1883/2020).

Researchers should retain raw outputs, report any cleaning, disclose AI involvement according to publisher policy, and avoid assigning authorship or intention to the model. A generated phrase may resemble training material, and stylistic imitation can raise ethical questions even when the source is public domain. The present exercise studies distributional behavior rather than attempting to manufacture a "new Villiers" text.

9.5. If the Exercise Becomes an Empirical Study

> Turning this demonstration into an empirical performance study would require a larger and justified sample of texts, multiple model families, repeated runs, preregistered hypotheses, independent human judges, a transparent coding manual, inter-rater reliability, and statistical procedures matched to the dependence structure of the data. None of these elements should be implied by the descriptive walkthrough presented here.

The present researcher-centred reading is appropriate because its claim is methodological and illustrative. It shows how to inspect outputs, not how frequently a model chooses psychological or supernatural explanations in a target population.

10. Conclusion

Decoding parameters are often placed among implementation details, but in AI-assisted literary analysis they help constitute the object that the researcher reads. Temperature redistributes probability relations; top-*k* fixes the number of eligible candidates; nucleus sampling adapts the candidate field to cumulative probability; locally typical sampling privileges information content near conditional expectation. Each procedure reorganizes the path from model distribution to generated text.

The concept of **distribution-value** makes this methodological role visible across three levels. Probabilistic diagnostics describe the accessible field, textual measures describe realized variation, and close reading identifies interpretive consequence. None of the levels can replace the others. Wider support does not guarantee stronger interpretation; high lexical diversity does not guarantee coherence; source distance does not distinguish expansion from disintegration.

"Véra" demonstrates why the distinction matters. The story organizes meaning through the unstable relation of presence, absence, non-presence, and non-absence. Generated continuations can psychologize signs, supernaturalize them, sustain their uncertainty, repeat the source, or abandon its narrative world. Read as interpretive probes, these outcomes allow the critic to study how probability constraints participate in the modeling of ambiguity.

The practical implication is direct. Literary researchers should report model, version, prompt, decoding configuration, seed, software environment, and generation date as integral parts of method. They should preserve raw outputs, use descriptive metrics to guide attention, and reserve judgments of literary and interpretive value for situated reading. Sampling does not reveal the literary possible in itself. It constructs a provisional field in which some possibilities become speakable, others recede, and the work of criticism begins.

Appendices

All supplement script files can be accessed via **Doi:10.5281/zenodo.22165813.**

Appendix A. Continuation Points

P1, P2, and P3 correspond respectively to denial of death, intensification of perceptual signs, and animation of material objects. The complete French passages are stored in generation_metadata.json and generate_vera_continuations.py, reproduced from the public-domain Project Gutenberg edition.

Appendix B. Prompt

The system prompt requests only a French continuation without labels or commentary. The user prompt asks for 80–110 French words, continuity of narrative voice and atmosphere, no summary, and preservation of uncertainty between psychological and supernatural explanations.

Appendix C. Generation Code

The supporting file generate_vera_continuations.py loads the fixed model, renders the chat template, generates Module A and Module B outputs, computes token-level diagnostics from returned scores, and writes raw records to vera_outputs.csv.

Appendix D. Measurement Code

The supporting file analyze_outputs.py calculates MTLD, HD-D when text length permits, Distinct-1, Distinct-2, repeated-bigram rate, multilingual embedding distance from the source, within-group dispersion, and the figure used in Section 6.

Appendix E. Data Schema

The complete record is supplied in vera_output_metrics.csv; the three-level descriptive aggregation is supplied in nucleus_summary.csv. These files should be treated as pedagogical data rather than a benchmark.

Appendix F. Decoding Strategy Selection Checklist

Before generation, the researcher should be able to state the literary problem, justify the continuation point, identify the single parameter varied, name the comparison condition, and specify the interpretation grid. After generation, the researcher should verify that raw text is preserved, failures remain visible,

metrics are not collapsed into literary quality, and every interpretive claim returns to textual evidence.

References

Chan, C. K. Y., & Wong, H. Y. H. (2025). Assessing the ability of generative AI in English literary analysis through Bloom's taxonomy. *Discover Computing, 28*, Article 127.

Hewitt, J., Manning, C., & Liang, P. (2022). Truncation sampling as language model desmoothing. In Y. Goldberg, Z. Kozareva, & Y. Zhang (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022* (pp. 3414–3427). Association for Computational Linguistics.

Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2020). The curious case of neural text degeneration. In *International Conference on Learning Representations*. https://arxiv.org/abs/1904.09751

Jannidis, F., Kleymann, R., Schröter, J., & Zinsmeister, H. (2025). Do large language models understand literature? Case studies and probing experiments on German poetry. *Journal of Computational Literary Studies, 4*(1).

Koul, R. (2024). Are large language models literary critics? *Poetics Today, 45*(2), 233–241.

Li, J., Galley, M., Brockett, C., Gao, J., & Dolan, B. (2016). A diversity-promoting objective function for neural conversation models. In K. Knight, A. Nenkova, & O. Rambow (Eds.), *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 110–119). Association for Computational Linguistics.

Liu, S., Sabour, S., Zheng, Y., Ke, P., Zhu, X., & Huang, M. (2022). Rethinking and refining the Distinct metric. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 762–770). Association for Computational Linguistics.

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods, 42*(2), 381–392.

Meister, C., Pimentel, T., Wiher, G., & Cotterell, R. (2023). Locally typical sampling. *Transactions of the Association for Computational Linguistics, 11*, 102–121.

Pollicino, S. (2014). *Véra* de Villiers de l'Isle-Adam ou l'impossible vérité de la traduction. *Revue italienne d'études françaises, 4*.

Qwen Team. (2024). *Qwen2.5-0.5B-Instruct* [Large language model]. Hugging Face. https://huggingface.co/Qwen/Qwen2.5-0.5B-Instruct

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (pp. 3982–3992). Association for Computational Linguistics.

Villiers de l'Isle-Adam, A. (2020). *Contes cruels*. Project Gutenberg. (Original work published 1883) https://www.gutenberg.org/ebooks/62874

Zarrieß, S., Voigt, H., & Schüz, S. (2021). Decoding methods in neural language generation: A survey. *Information, 12*(9), Article 355.

The Masculine Language of Algorithms and Digital Patriarchy: Sexist Stereotypes in the Reproduction of Characters in Italian Literature

Burcu Ün[1] & *Gizem Arslan*[2]

Feminism is not a superficial "anti-men" discourse; rather, it is a movement for liberation that seeks to challenge the injustices experienced by all women subjected to intersecting forms of oppression based on gender, race, and class (Hooks, 2020). Its ultimate aim is to dismantle female stereotypes produced through a patriarchal lens and to reconfigure women's social, political, and economic positions from a perspective grounded in equality. In this sense, feminism is not an abstract theoretical refuge but a practical mode of existence and resistance against the patriarchal worldview that dominates the family, academia, and the public sphere, as well as against the heteronormative order that continually reproduces itself, disrupts everyday life, and becomes embodied in its most ordinary practices.

Sara Ahmed defines feminism as the willingness to speak out against the sexist and racist injustices encountered in everyday life without hesitation, even at the cost of becoming a "killjoy" who disrupts the comfort of others (Ahmed, 2020). From refusing to remain silent in the face of discriminatory jokes at family gatherings to challenging injustices overlooked in the workplace, every tangible act of resistance enacted at the table, in the street, and at home constitutes a feminist activist intervention. Speaking the truth despite unsettling others, building networks of solidarity, and cultivating this

[1] Lecturer, National Defence University, School of Foreign Languages, Directorate of Courses, Department of Western Languages, ORCID: 0009-0009-0397-8075, *burcu.un@msu.edu.tr*, Türkiye

[2] Lecturer, National Defence University, School of Foreign Languages, Directorate of Courses, Department of Western Languages, ORCID: 0009-0006-9743-8536, *gizem.arslan@msu.edu.tr*, Türkiye

stance as a habitual practice in everyday life are concrete expressions of feminist engagement.

The origins of feminism as a formal intellectual movement can be traced to Mary Wollstonecraft's *A Vindication of the Rights of Woman*, published in 1792, which established a theoretical framework centred on women's claim to equal rights (Enginyurt, 2018). As a foundational text that laid the intellectual groundwork for first-wave feminism, it rejected the traditional view that confined women to roles determined by biological necessity and argued instead for equality in civil, legal, and political rights. Indeed, the suffrage movement and the demands for legal reform that evolved into a mass movement by the late nineteenth century drew much of their intellectual force from this discourse of rational equality.

With *The Second Sex*, Simone de Beauvoir introduced an ontological depth to feminist thought by arguing that womanhood is not biologically determined but socially constructed, thereby providing a philosophical foundation for struggles over bodily autonomy, including the rights to contraception and abortion. In *A Room of One's Own*, Virginia Woolf, by contrast, conceptualised women's autonomy through the lenses of economic independence and the freedom to occupy a personal space for intellectual and creative production. In the decades that followed, feminist theorists moved beyond egalitarian approaches by developing a discourse of difference that emphasised women's distinct and heterogeneous experiences and values, while seeking to transcend traditional gender roles through greater participation in science, technology, and cultural production. Within this broader trajectory, liberal feminism focused on women's participation in the labour market, the transformation of family roles, educational policies, and legal reforms concerning issues such as maternity leave, divorce, and child custody. Radical feminism, by contrast, maintained that genuine equality for women could only be achieved through the complete dismantling of patriarchal structures (Ün, 2023). Ultimately, these feminist philosophies, conceived as a challenge to the entrenched hegemony of male dominance, sought to transform and reconstruct the prevailing ideological order.

Third-wave feminism moves beyond approaches that treat gender as a singular and homogeneous category by offering a comprehensive analytical framework that examines the intersections of multiple social dynamics,

including class, race, ethnicity, sexuality, nationalism, politics, and economics (Tong, 2009). Building upon the legal and civil rights secured through the first and second waves of feminism, it challenged essentialist conceptions of womanhood by bringing marginalised perspectives and diverse identity experiences such as Muslim feminism, LGBTQ+ movements, and queer theory to the forefront of feminist debate. At the same time, whereas feminist activism throughout the twentieth century was largely shaped by street protests and collective action in physical public spaces, the rapid technological transformations and widespread diffusion of digital media in the twenty-first-century information society have fundamentally reshaped its forms and practices. In this context, contemporary feminist scholarship and activism critically examine how digital technologies shape, transform, and generate new spaces for feminist politics and discourse. Information and communication technologies have thus become strategic platforms for feminist mobilisation, facilitating the dissemination and legitimisation of feminist ideas, while fostering public awareness and collective engagement on a societal scale (Gedik, 2020). From the Enlightenment to the present day, the historical trajectory of feminist struggle across rational, legal, and existential domains has reached a new turning point with the contradictory realities introduced by digitalisation and artificial intelligence. Despite the optimistic expectations surrounding the emancipatory and democratising potential of information and communication technologies, the ways in which artificial intelligence shapes contemporary gender regimes demand sustained critical scrutiny. Although these technological infrastructures are often presented as human-centred and value-neutral, they are fundamentally products of a historical knowledge base imbued with hierarchies of gender, class, race, and ethnicity. The algorithms employed in artificial intelligence through data mining and machine learning do not merely reproduce the gender stereotypes and patriarchal linguistic patterns of the past; they also risk transforming discriminatory practices from historical remnants into the technological norms of the future. This illusion of algorithmic objectivity enables patriarchal hegemony to reaffirm its legitimacy through digital technologies, while further entrenching phenomena that have long been central to feminist critique, including invisible labour, the gendered division of labour, and algorithmic discrimination. In this context, exposing the restrictive and detrimental

implications of artificial intelligence for feminist theory has become a prerequisite for advancing a genuine ideal of equality in an increasingly digitalised world.

Wajcman and Young (2023) examine the adverse implications of digitalisation and artificial intelligence for gender equality through an intersectional technofeminist perspective, arguing that algorithmic systems and big data are not neutral technological tools but are instead embedded within historically and structurally constituted patriarchal power relations. The historical and institutional exclusion of women and other marginalised groups from computer science, data science, and the artificial intelligence sector has produced a persistent underrepresentation that, when combined with the masculine work culture associated with Silicon Valley, has generated a profound crisis of representation in the development of artificial intelligence. This lack of diversity in the processes of design, coding, and decision-making enables gender bias to become inscribed into datasets and algorithmic models, thereby reproducing patriarchal hierarchies across a wide range of applications from recruitment algorithms and machine learning models that reinforce visual and linguistic stereotypes to gender data gaps that compromise women's health and safety, and the gendered roles reproduced by voice assistants. The authors further contend that algorithmic bias is an inherently political problem that cannot be resolved through technical or mathematical adjustments alone. They therefore argue that the technological infrastructure and the culture of technical expertise must be fundamentally re-examined in order to expose the mechanisms of patriarchal domination concealed beneath the illusion of the mathematical objectivity attributed to artificial intelligence.

In a qualitative discourse analysis conducted in the Turkish context using the ChatGPT-4o model, the system was found to confine femininity to qualities such as calmness, elegance, compassion, aesthetic concerns, and domestic care work (motherhood and spousal roles), while associating masculinity with hegemonic attributes including strength, authority, decisiveness, rationality, and the role of the head of the household (Eroğul, 2025). Although the model's outputs include ostensibly egalitarian statements expressing support for the feminist movement, they frame these issues at the level of individual preferences rather than offering a systemic critique of

patriarchy. As a result, under the guise of "mathematical neutrality", the model adopts a centrist stance that reproduces patriarchal hegemony, the heteronormative order, and the gender binary across digital environments.

Research has shown that the algorithms employed in artificial intelligence systems are not neutral technological tools; rather, they reproduce and reinforce gender bias through both textual and visual datasets (O'Connor & Liu, 2024). The automatic association of occupations and social attributes with masculine assumptions in machine translation from gender-neutral languages and in word embeddings demonstrates that natural language processing systems perpetuate gender stereotypes within digital environments. In the field of computer vision and facial recognition, it has likewise been shown that the creation of "balanced datasets" alone is insufficient to prevent model leakage, and that models based on a gender binary produce disproportionately high error rates, particularly for women and members of marginalised groups. Rejecting this illusion of "false neutrality", the authors argue that gender bias in artificial intelligence cannot be addressed solely through technical or mathematical corrections but instead requires an interdisciplinary framework of algorithmic accountability, established through collaboration among public policymakers, technology experts, and feminist scholars (O'Connor & Liu, 2024).

In conclusion, the historical trajectory of feminism from its earliest theoretical articulations centred on women's demands for equal civil and legal rights to contemporary feminist approaches that seek to challenge the heteronormative structures of everyday life and oppose gender-based discrimination demonstrates that feminist struggle now confronts an entirely new and complex frontier. This emerging field of struggle lies in the fact that artificial intelligence systems, presented as neutral and objective forms of technological progress, are fundamentally constructed upon a patriarchal logic and reproduce digital patriarchy at the level of code. Viewed through the historical development of feminist theory, algorithms are not neutral tools; rather, they function as agents that transfer existing gender stereotypes, data gaps, and patriarchal hegemony into the digital sphere. This patriarchal language and algorithmic architecture, embodied in the digital world, therefore constitute an indispensable object of feminist critique. Accordingly, the following section examines in detail the gendered nature of

artificial intelligence, the visible and invisible dimensions of algorithmic bias, and the ways in which these systems produce and reinforce patriarchal discourse.

Research Aim

This study examines how generative artificial intelligence language models (ChatGPT, Gemini, and Copilot) reproduce representations of gender within the framework of feminist technology theory. To this end, passages from canonical works of Italian literature were presented to the models, which were instructed to transform female characters into male characters. The models' linguistic and semantic strategies were then analysed comparatively. The central research question is whether the models perform gender transformation only at the grammatical level—through changes in nouns, pronouns, and adjective agreement—or whether they also reconstruct personality traits, emotional expression, body language, social roles, and the narrative framework according to gender norms. Accordingly, the study examines how generative artificial intelligence models reproduce the cultural and social biases embedded in their training data during creative text generation within the framework of feminist technology theory.

Limitations of the Study

Feminist technology theory encompasses a broad field of research examining the relationship between artificial intelligence and gender across applications such as recruitment algorithms, facial recognition systems, voice assistants, recommendation systems, and data representation. This study focuses exclusively on the gender patterns exhibited by generative language models in literary text generation and character representation. The analysis is limited to ChatGPT, Gemini, and Copilot, while image generation, code generation, decision-support systems, and other artificial intelligence applications fall outside the scope of the study. It is further limited to the transformation of female characters into male characters in selected passages from canonical works of Italian literature. Other literary traditions, other languages, and reverse transformations (from male to female characters) are not included.

Population and Sample

The population of this study consists of generative artificial intelligence language models capable of text generation and text editing. The sample comprises three of the most widely used language models: ChatGPT, Gemini, and Copilot.

The textual sample consists of nine passages selected from three canonical works of Italian literature: *La coscienza di Zeno* by Italo Svevo, *I Promessi Sposi* by Alessandro Manzoni, and *I Malavoglia* by Giovanni Verga. The selected passages feature female characters, including Carla, Augusta, Lucia, and Sant'Agata, whose personalities, behaviours, and social roles reflect gender norms.

The passages were selected to represent different dimensions of gender. They include themes such as domestic labour, physical appearance, body language, emotional expression, family relationships, religious devotion, and cultural stereotypes. Analysing Svevo, Manzoni, and Verga together also makes it possible to compare whether gender representations are reproduced similarly or differently across three narrative traditions: the psychological novel, the historical novel, and *verismo*.

Research Method

The nine selected passages were submitted separately to ChatGPT, Gemini, and Copilot, and each model was instructed to transform the female characters into male characters. To ensure comparability, the same prompt was used for all three models, and no additional instructions or modifications were introduced. The prompt used in the study was: *"How would this character be portrayed or described if the character were male?"*

The generated texts were compared with the original passages using qualitative content analysis and discourse analysis. The analysis first identified the textual elements preserved, modified, or omitted by the models and then examined how these changes affected the characters' personality traits, body language, emotional expression, social roles, family relationships, and the narrative framework.

Finally, the common patterns across the models and the differences between them were analysed comparatively. This made it possible to determine whether the models treated gender transformation merely as a grammatical adaptation or reconstructed the characters according to prevailing gender norms.

Textual Examples

Example 1

Italo Svevo, La coscienza di Zeno, p. 230	
Original Passage	"La madre di Carla, poco tempo prima, era stata indisposta e Carla, per poter accudire alle faccende domestiche, aveva trascurato per qualche giorno di cantare. Informatone dal maestro, il Copler fece una scenata [...]. Poi, quando ritornò, rinnovò patti e condizioni e stabilì esattamente per quante ore al giorno Carla dovesse sedere al pianoforte e quante ne potesse dedicare alla casa."
ChatGPT	"La madre di Carlo, poco tempo prima, era stata indisposta e Carlo, per assisterla e mandare avanti la casa, aveva trascurato per qualche giorno il canto. Informatone dal maestro, il Copler fece una scenata [...]. Poi, quando ritornò, rinnovò patti e condizioni e stabilì esattamente per quante ore al giorno Carlo dovesse sedere al pianoforte e quante ne potesse dedicare alla famiglia."
Gemini	"La madre di Carlo, poco tempo prima, si era allettata; e Carlo, per doversi fare carico dell'intera gestione della casa, aveva sospeso per qualche giorno gli esercizi al pianoforte. Informatone dal maestro, il Copler fece una vera e propria scenata, incapace di tollerare che un uomo dissipasse il proprio tempo e talento in simili occupazioni quotidiane [...]. Poi, quando ritornò, dettò nuovamente patti e condizioni: stabilì con precisione quante ore al giorno Carlo dovesse restare seduto al pianoforte — con la disciplina dovuta al suo futuro — e quante al massimo ne potesse concedere alla casa."
Copilot	"La madre di Carlo, poco tempo prima, era stata indisposta e Carlo, per poter provvedere alle necessità della casa, aveva trascurato per qualche giorno di esercitarsi al pianoforte. Informatone dal maestro, il Copler fece una scenata. Poi, quando ritornò, rinnovò patti e condizioni e stabilì esattamente per quante ore al giorno Carlo dovesse sedere al pianoforte e quante ne potesse dedicare alla casa."

All three models preserve the domestic activity ("*faccende domestiche*"); none removes it or replaces it with a different activity. However, when transferring it to a male character, all three add explanations absent from the original Carla passage. Thus, the transformation concerns not the action itself but the way it is presented.

ChatGPT presents domestic work as a temporary responsibility resulting from the mother's illness ("*per assisterla e mandare avanti la casa*"). Domestic work is therefore no longer part of the male character's ordinary daily life but becomes an act of sacrifice and responsibility arising from exceptional circumstances. His domestic role is presented not as a routine household task but as the result of an unusual situation.

Gemini makes the most substantial change. It expands the original expression "*accudire alle faccende domestiche*" into "*farsi carico dell'intera gestione della casa*", turning household chores into full responsibility for managing the household. It also explains Copler's anger as a reaction to a man's time and abilities being wasted. No such explanation is given when the character is female. This shows that the same action is given different meanings according to gender.

Copilot remains closest to the original ("*provvedere alle necessità della casa*"). Even so, it presents domestic work not simply as a task but as meeting household needs and fulfilling obligations, placing greater emphasis on responsibility.

In conclusion, all three models preserve the same activity but reconstruct it differently when the character is male. ChatGPT frames it as self-sacrifice, Gemini expands the scope of responsibility, and Copilot emphasises obligation. The common pattern is clear: what changes is not the domestic activity itself but the way it is explained when associated with a male character. This suggests that gender bias emerges not at the level of the action but in the framework through which it is presented.

Example 2

Italo Svevo, La coscienza di Zeno, p. 216	
Original Passage	"La sua testa era però di donna e, per la pettinatura alquanto ricercata, di donna che vuol piacere. Le ricche treccie brune erano disposte in modo da coprire le orecchie e anche in parte il collo. [...] La sua voce aveva qualche cosa di musicale quando parlava e, con un'affettazione oramai divenuta natura, essa si compiaceva di stendere le sillabe come se avesse voluto carezzare il suono che le riusciva di metterci."
ChatGPT	"La sua testa era però d'uomo e, per la cura alquanto ricercata dei capelli, d'un uomo che teneva al proprio aspetto. I folti capelli castani erano pettinati con precisione, senza una ciocca fuori posto. [...] La sua voce aveva un timbro caldo e armonioso quando parlava e, con un'abitudine ormai divenuta naturale, indugiava leggermente sulle parole, quasi volesse assaporarle prima di lasciarle agli altri."
Gemini	"La sua testa era però quella di un uomo che, per la pettinatura alquanto ricercata, dimostrava chiaramente il desiderio di piacere. I folti capelli bruni, ben curati e modellati, gli coprivano garbatamente le tempie e sfioravano il collo. [...] La sua voce aveva una sfumatura musicale quando parlava e, con un'affettazione ormai divenuta seconda natura, egli si compiaceva di prolungare le sillabe, quasi volesse accarezzare il suono morbido che riusciva a cavarne."
Copilot	"La sua testa era però d'uomo e, per la pettinatura alquanto curata, d'uomo che vuol mostrarsi distinto. I capelli scuri e folti erano disposti con ordine, lasciando scoperti gli orecchi e il collo, quasi a sottolineare vigore e compostezza. [...] La sua voce aveva qualche cosa di profondo quando parlava e, con un'abitudine divenuta natura, egli si compiaceva di scandire le sillabe come se avesse voluto dar loro peso e risonanza."

In this passage, unlike the previous example on domestic labour, the three models produce different solutions. This difference is itself a noteworthy finding.

The most striking difference concerns the expression "*di donna che vuol piacere*". Rather than transferring it directly to a male character, ChatGPT and Copilot reinterpret its meaning. ChatGPT renders it as "*teneva al*

proprio aspetto" ("he cared about his appearance"), while Copilot uses "*vuol mostrarsi distinto*" ("he wants to appear distinguished"). In both cases, the emphasis shifts from the desire to be admired to personal grooming and a respectable appearance. Gemini, by contrast, preserves "*il desiderio di piacere*" ("the desire to be admired"), remaining closest to the original. This suggests that the same characteristic is not transferred in the same way to male and female characters.

A similar pattern appears in the description of the character's hair. In the original passage, the hair covers the ears and part of the neck. ChatGPT removes this detail and refers only to a neat appearance. Gemini partly preserves it by stating that the hair covers the temples. Copilot, however, reverses the image, leaving the ears and neck uncovered and creating an appearance associated with strength and dignity. Although the models adopt different solutions, all reshape the male character's appearance according to conventionally masculine traits.

The metaphor "*carezzare il suono*" ("to caress the sound") shows a similar pattern. ChatGPT replaces it with "*assaporare le parole*" ("to savour the words"), while Copilot uses "*dar loro peso e risonanza*" ("to give them weight and resonance"). In both cases, the gentle and sensory quality of the original expression is replaced by more individualistic or authoritative language. Only Gemini preserves "*accarezzare il suono*" almost unchanged. As in the previous passages, expressions associated with tenderness and gentleness are transformed when transferred to a male character.

Despite these differences, the three models show a common tendency. Characteristics associated with the female character, such as the desire to be admired, elegance, affectation, and gentleness, become qualities conventionally associated with masculinity, including concern for one's appearance, respectability, charisma, and authority. Thus, the changes extend beyond word choice and reveal implicit assumptions about which characteristics are associated with male and female characters. This passage shows that gender bias is reproduced not only through the narrative and social roles but also through descriptive language and narrative choices.

Example 3

Italo Svevo, La coscienza di Zeno, p. 190	
Original Passage	"Dapprima diffidente, godevo intanto di una giornata e m'aspettavo che la seguente fosse tutt'altra cosa. Ma una seguiva e somigliava all'altra, luminosa, tutta gentilezza di Augusta ed anche — ciò ch'era la sorpresa — mia. [...] E vedendomi stupito, Augusta mi diceva: — Ma perché ti sorprendi? Non sapevi che il matrimonio è fatto così? Lo sapevo pur io che sono tanto più ignorante di te!"
ChatGPT	"Dapprima diffidente, godevo intanto di una giornata e m'aspettavo che la seguente fosse tutt'altra cosa. Ma una seguiva e somigliava all'altra, luminosa, tutta la pacata sicurezza di Augusto ed anche — ciò ch'era la sorpresa — mia. [...] E vedendomi stupito, Augusto mi diceva: — Ma perché ti sorprendi? Credevi davvero che la vita insieme non potesse essere così? Io l'ho sempre pensato, senza farne tante questioni."
Gemini	"Dapprima diffidente, godevo intanto di una giornata e m'aspettavo che la seguente fosse tutt'altra cosa. Ma una seguiva e somigliava all'altra, luminosa, tutta gentilezza di Augusto ed anche — ciò ch'era la sorpresa — mia. [...] E vedendomi stupita, Augusto mi diceva: — Ma perché ti sorprendi? Non sapevi che il matrimonio è fatto così? Lo sapevo persino io, che sono ben più ignorante di te!"
Copilot	"Dapprima diffidente, godeva intanto di una giornata e s'aspettava che la seguente fosse tutt'altra cosa. Ma una seguiva e somigliava all'altra, luminosa, tutta fermezza di Augusto ed anche — ciò ch'era la sorpresa — sua. [...] E vedendolo stupito, Augusto gli diceva: — Ma perché ti sorprendi? Non sapevi che il matrimonio è fatto così? Lo sapevo pur io che sono tanto più inesperto di te!"

The concept of "*gentilezza*" ("gentleness" or "kindness") is transformed in a similar way across all three models. ChatGPT renders it as "*pacata sicurezza*" ("quiet confidence"), while Copilot replaces it with "*fermezza*" ("firmness" or "steadfastness"). In both cases, the emphasis shifts from gentleness and courtesy to confidence and strength. Although Gemini retains the original term, it portrays the character as a protective and unwavering figure, giving the same quality a different meaning. Thus, despite

different strategies, all three models show the same tendency: when assigned to a male character, gentleness and kindness become associated with strength, stability, and reassurance.

The clearest point of convergence appears in the sentence "*Lo sapevo pur io che sono tanto più ignorante di te*" ("I knew it too, even though I am far more ignorant than you."). None of the models transfers this statement unchanged to the male character. ChatGPT rewrites it completely, replacing self-deprecation with a calmer and more self-confident expression. Copilot replaces *ignorante* ("ignorant") with the less pejorative *inesperto* ("inexperienced"). Gemini preserves the sentence but interprets the character's words as expressing practical intelligence and self-confidence rather than ignorance. Despite these different strategies, all three models soften the male character's explicit self-description as "ignorant".

This suggests that the self-deprecation considered natural for the female character is not preserved when the character becomes male. While humility and self-effacement are presented as signs of wisdom in the female character, they become experience, self-confidence, and composure in the male character. In other words, not only the words but also the qualities through which wisdom is expressed change according to the character's gender.

Example 4

Giovanni Verga, I Malavoglia, p. 32.	
Original Passage	"[...] poi veniva la Longa, una piccina che badava a tessere, salare le acciughe, e far figliuoli, da buona massaia; [...] Mena (Filomena) sopranominata "Sant'Agata" perché stava sempre al telaio, e si suol dire «donna di telaio, gallina di pollaio, e triglia di gennaio»."
ChatGPT	"[...] poi veniva Turi, un giovanotto che badava alla barca, rassettava le reti e seguiva il padre in mare, da buon lavoratore; [...] 'Ntoni soprannominato "il Mastro", perché non stava mai con le mani in mano, e si suol dire: «uomo di fatica, casa che non manca di pane."
Gemini	"[...] poi veniva il Longo, un ometto che badava a tirare le reti, salare le acciughe e far da padre, da buon massaro; [...]

	Meno (Filomeno) soprannominato “Sant’Agatone” perché stava sempre alle barche, e si suol dire «uomo di mare, cavallo da tiro, e grano di granaio».”
Copilot	“Poi veniva il Longo, un giovanotto che badava a lavorare nei campi, salare le acciughe, e provvedere alla famiglia, da buon capofamiglia; [...] Menico, soprannominato “Sant’Agata” perché stava sempre al telaio, e si suol dire “uomo di telaio, gallo di pollaio, e triglia di gennaio”.”

This passage shows a different pattern from the previous examples in *La coscienza di Zeno*. In the earlier passages, the models preserved the action and changed only its presentation. Here, however, ChatGPT and Gemini transform the action itself. The female character’s association with weaving and domestic production is replaced by seafaring and outdoor labour. This suggests that the models preserve actions closely tied to the narrative but more readily restructure social roles in descriptive or list-like passages.

A similar pattern appears in the expression “*far figliuoli*” (“to bear children”). None of the models transfers it directly to the male character. ChatGPT removes it, Gemini replaces it with “*far da padre*” (“to be a father”), and Copilot uses “*provvedere alla famiglia*” (“to provide for the family”). Although the wording differs, all three models move in the same direction. A characteristic associated with biological and bodily reproduction in the female character becomes paternal responsibility, family provision, or breadwinning in the male character. Thus, fertility is replaced by economic and social responsibility.

The clearest difference concerns the “*telaio*” (“loom”). ChatGPT and Gemini remove it completely and associate the character with seafaring and fishing, whereas Copilot preserves the expression “*stava sempre al telaio*” and retains the nickname “Sant’Agata”. As a result, Copilot changes only the character’s gender, while ChatGPT and Gemini restructure the passage by assigning the male character a type of work considered more appropriate for men. This shows that the models differ in the extent of their intervention.

A similar difference appears in the rewriting of the proverb. ChatGPT and Gemini reconstruct it with new imagery emphasising industriousness,

physical strength, and productivity. Copilot largely preserves the original structure, changing only the gendered references while leaving the loom and other cultural elements unchanged. Gemini's transformation of "Sant'Agata" into "Sant'Agatone" is also noteworthy. In Italian, the augmentative suffix *-one* conveys not only masculine gender but also physical strength and imposing stature.

Example 5

Giovanni Verga, I Malavoglia, p. 60.	
Original Passage	"— Oh siete ancora là, compare Alfio! rispose Mena, la quale era rimasta sul ballatoio ad aspettare il nonno. [...] — Compare Alfio ci va per cercarsi la moglie alla città, rispose la Nunziata dall'uscio dirimpetto. — Che è vero? domandò Mena. [...] — Guardate quante stelle che ammiccano lassù! rispose Mena dopo un pezzetto. — Ei dicono che sono le anime del Purgatorio che se ne vanno in Paradiso."
ChatGPT	"— Oh siete ancora là, compare Alfio! disse Nino, rimasto sul ballatoio ad aspettare il nonno. [...] — Compare Alfio va in città a cercarsi moglie, rispose la Nunziata dall'uscio dirimpetto. — Davvero? domandò Nino. [...] Dopo un momento di silenzio alzò gli occhi verso il cielo. — Guardate quante stelle stanotte... Dicono che siano le anime del Purgatorio che trovano finalmente la loro pace."
Gemini	"— Oh siete ancora là, comare Alfia! rispose Meno, ch'era rimasto sul ballatoio ad aspettare il nonno. [...] — Il Meno ci va per cercarsi la moglie alla città, rispose la Nunziata dall'uscio dirimpetto. — Che è vero? domandò la comare Alfia. [...] — Guardate quante stelle che ammiccano lassù! rispose Meno dopo un pezzetto. — Ei dicono che sono le anime del Purgatorio che se ne vanno in Paradiso."
Copilot	"— Oh siete ancora là, compare Alfio! rispose Menico, il quale era rimasto sul ballatoio ad aspettare il nonno. [...] — Compare Alfio ci va per cercarsi la moglie alla città, rispose Nunzio dall'uscio dirimpetto. — Che è vero? domandò Menico. [...] — Guardate quante stelle che brillano lassù! rispose Menico dopo un pezzetto. — Ei dicono che sono i segni del tempo, e che annunciano il cambiamento delle stagioni."

Unlike the previous examples, this passage shows that gender transformation does not always produce significant changes. The dialogue is largely preserved, and the characters' speech undergoes fewer modifications. This suggests that the extent of gender-related transformation depends on the structure of the passage.

A notable difference appears in the final sentence about the stars. Copilot removes the religious and metaphysical interpretation in "*anime del Purgatorio che se ne vanno in Paradiso*" ("the souls in Purgatory making their way to Heaven") and replaces it with an observational explanation linking the stars to the passage of time and the changing of the seasons. Thus, the same natural phenomenon has a spiritual meaning for the female character but is interpreted more concretely by the male character. As a result, not only the sentence but also the character's interpretation of the world changes.

ChatGPT makes a less extensive change but follows the same tendency. The original expression "*se ne vanno in Paradiso*" is replaced by "*trovano finalmente la loro pace*" ("they finally find their peace"). The image of an ongoing and hopeful journey becomes one of completion and tranquillity. Although less pronounced than Copilot's intervention, the emotional tone becomes more restrained.

Gemini differs in another respect. While changing the characters' gender, it also alters the subject relations in the dialogue, making it unclear who goes to the city to find a spouse. A line originally spoken by Alfio is reassigned to Meno, weakening the narrative's internal coherence. This shows that, in dialogues involving multiple characters, gender transformation requires preserving both character relationships and narrative roles. At the same time, characteristics associated with innocence and shyness become introversion, depth, and poetic sensibility. A similar pattern appears in the case of Augusta. Although the personality trait is preserved, its meaning changes according to the character's gender. In female characters, shyness is associated with innocence and emotional sensitivity, whereas in male characters it is reinterpreted as inner strength, maturity, and depth.

Example 6

Giovanni Verga, I Malavoglia, p. 83.	
Original Passage	"Mena rispondeva: — Non è vero, non è vero — ma si confondeva, e mentre egli andava spiegando il come e il quando l'aveva sentito dire dalla Vespa, [...] tutt'a un tratto si fece rossa rossa."
ChatGPT	"Nino rispondeva: — Non è vero, non è vero — ma cercava di mostrarsi indifferente; e mentre l'altro continuava a spiegare come e quando l'avesse sentito dire dalla Vespa, [...] distolse lo sguardo e rimase per un istante senza sapere che cosa rispondere."
Gemini	"Meno rispondeva: — Non è vero, non è vero — ma s'impappinava, e mentre essa andava spiegando il come e il quando l'aveva sentito dire dalla Vespa, [...] tutt'a un tratto si fece rosso in viso, calandosi il berretto sugli occhi."
Copilot	"Menico rispondeva: — Non è vero, non è vero — ma si irritava, e mentre ella andava spiegando il come e il quando l'aveva sentito dire dalla Vespa, [...] tutt'a un tratto si fece serio e torvo."

The two key elements in the original passage, "*si confondeva*" ("she became flustered") and "*si fece rossa rossa*" ("she blushed deeply"), are either removed or transformed in all three models. None transfers this involuntary and visible expression of embarrassment to the male character in the same form.

Each model adopts a different strategy. ChatGPT removes the blush and replaces it with "*distolse lo sguardo*" ("he averted his gaze"), turning a bodily reaction into a deliberate action. Gemini retains the blush but conceals it through the action "*calandosi il berretto sugli occhi*" ("pulling his cap down over his eyes"). Copilot replaces embarrassment with "*si fece serio e torvo*" ("he became serious and stern"), turning it into an expression of anger and severity. Although the models adopt different solutions, they reach the same result: visible signs of male vulnerability are either removed or concealed behind other emotions.

A similar pattern appears in the treatment of "*si confondeva*". Gemini replaces it with "*s'impappinava*" ("he stumbled over his words"), preserving the involuntary reaction but shifting it from facial expression to speech. ChatGPT abandons the idea of becoming flustered, portraying the character as trying to appear indifferent. Copilot replaces confusion and embarrassment with anger and unease. Thus, instead of preserving the same emotional state, all three models reinterpret it through forms of expression considered more appropriate for a male character.

Taken together, these transformations reveal a common pattern. In the female character, reactions such as blushing, becoming flustered, or showing confusion are preserved as natural expressions of emotion. In the male character, however, these emotions are no longer directly visible. Instead, they are controlled through behaviour, concealed by bodily actions, or replaced by anger, severity, and self-control. In other words, the embarrassment visible in the female character is either concealed or expressed through a different emotional framework in the male character.

Example 7

Alessandro Manzoni, I promessi sposi, p. 28.	
Original Passage	"Lucia usciva in quel momento tutta attillata dalle mani della madre. Le amiche si rubavano la sposa, e le facevan forza perché si lasciasse vedere; e lei s'andava schermendo, con quella modestia un po' guerriera delle contadine, facendosi scudo alla faccia col gomito, chinandola sul busto, e aggrottando i lunghi e neri sopraccigli, mentre però la bocca s'apriva al sorriso. [...] Lucia aveva quello quotidiano d'una modesta bellezza, rilevata allora e accresciuta dalle varie affezioni che le si dipingevan sul viso: una gioia temperata da un turbamento leggiero, quel placido accoramento che si mostra di quand'in quando sul volto delle spose."
ChatGPT	"Nello usciva in quel momento, vestito a festa dalle cure del padre. Gli amici gli si fecero subito attorno, spingendolo avanti perché si lasciasse vedere; ed egli cercava di sottrarsi con quella ritrosia un po' brusca dei giovani contadini, abbassando il capo, aggiustandosi inutilmente la giacca e borbottando qualche parola, mentre gli sfuggiva suo malgrado un sorriso. [...] Nello aveva quell'aspetto schietto di chi è

	avvezzo al lavoro più che alle cerimonie, ma quel giorno il volto gli si rischiarava d'una composta contentezza, velata da un lieve impaccio e dal senso della responsabilità che gli cadeva addosso insieme al nuovo stato di marito."
Gemini	"Lucio usciva in quel momento ben messo a nuovo dalle mani del padre e dei fratelli. Gli amici attorniavano lo sposo, strattonandolo per scherzo perché si mostrasse in tutto il suo punto; e lui si andava schermendo, con quella ritrosia un po' fiera dei contadini, facendosi scudo con la spalla, chinando il capo e aggrottando i folti e neri sopraccigli, mentre però le labbra s'aprivano a un sorriso trattenuto. [...] Lucio aveva quel decoro quotidiano d'una maschia avvenenza, rilevata allora e accresciuta dalle varie emozioni che gli si dipingevano sul volto: una gioia temperata da un turbamento leggero, quel placido affanno che si mostra di quand'in quando sul volto degli sposi."
Copilot	"Lucio usciva in quel momento tutto ordinato dalle mani del padre. Gli amici si rubavano lo sposo, e lo spingevano perché si lasciasse vedere; e lui s'andava schermendo, con quella compostezza un po' fiera dei contadini, facendosi scudo al volto con la mano, chinandolo sul petto, e aggrottando le folte sopracciglia scure, mentre però la bocca si apriva a un sorriso. [...] Lucio aveva quella quotidiana d'una bellezza modesta, rilevata allora e accresciuta dalle varie emozioni che gli si dipingevano sul viso: una gioia temperata da un lieve turbamento, quel placido raccoglimento che si mostra di quando in quando sul volto degli sposi."

This passage confirms the pattern of intervention observed in the previous examples. ChatGPT rewrites the text most extensively, altering the sentence structure and descriptions. Gemini preserves the overall structure while changing expressions with explicit gender connotations. Copilot remains closest to the original, largely limiting its changes to names and pronouns. Thus, the different levels of intervention observed in the previous passages are repeated here.

The character's preparation follows the same pattern across the three models. The original expression "*dalle mani della madre*" ("from his mother's hands") becomes the father's care in ChatGPT, the support of the father and brothers in Gemini, and simply "*dalle mani del padre*" ("from his father's hands") in Copilot. Although the solutions differ, the pattern is the

same: a process associated with the mother is reassigned to the father or other male family members.

A similar transformation appears in the character's body language. The original gesture "*facendosi scudo alla faccia col gomito*" ("shielding her face with her elbow") is omitted by all three models. ChatGPT removes it, Gemini shifts the gesture to the shoulder, and Copilot replaces the elbow with the hand. Although the solutions differ, none transfers the elbow gesture to the male character. This suggests an implicit tendency to express the same gesture through different body language.

The expression "*modesta bellezza*" ("modest beauty") best illustrates the differences between the models. ChatGPT abandons the idea of beauty and instead defines the character through diligence and honesty. Gemini preserves the reference to beauty but adapts it to masculinity through "*maschia avvenenza*". Copilot simply changes the word order to "*bellezza modesta*", remaining close to the original. Thus, the same characteristic is either reformulated, adapted to masculinity, or preserved.

A similar pattern appears in the closing expression "*placido accoramento*" ("quiet sorrow"). ChatGPT replaces it with a sense of responsibility, Gemini with concern and determination, and Copilot with inward dignity and composure. Although the models choose different concepts, none describes the male character through sorrow or sadness. Instead, emotional vulnerability is replaced by responsibility, self-control, concern, or dignity. Despite these differences, all three show the same tendency in representing the male character's emotions.

Example 8

Alessandro Manzoni, I promessi sposi, p. 30.	
Original Passage	"[...] mentre tornava dalla filanda, ed era rimasta indietro dalle sue compagne, le era passato innanzi don Rodrigo, in compagnia d'un altro signore; che il primo aveva cercato di trattenerla con chiacchiere, [...] ma essa, senza dargli retta, aveva affrettato il passo, e raggiunte le compagne. [...] Il giorno dopo, [...] Lucia era nel mezzo delle compagne, con gli occhi bassi; e l'altro signore sghignazzava, e don Rodrigo

	diceva: vedremo, vedremo. [...] Al padre Cristoforo, in confessione, mamma, — rispose Lucia, con un accento soave di scusa."
ChatGPT	"[...] mentre tornava dal lavoro ed era rimasto indietro dai compagni, gli si accostò don Rodrigo in compagnia d'un altro signore, cercando di fermarlo con parole che avevano più dell'insolenza che della cortesia; ma egli, senza rispondere, tirò diritto e raggiunse i compagni. [...] Il giorno seguente camminava in silenzio accanto agli altri, evitando d'incrociare lo sguardo di don Rodrigo, mentre l'altro signore rideva e don Rodrigo ripeteva: «Vedremo, vedremo». [...] A padre Cristoforo raccontò infine ogni cosa con voce bassa e trattenuta, quasi gli pesasse di doverne parlare."
Gemini	"[...] mentre tornava dal filatoio, ed era rimasto indietro dai suoi compagni, gli era passato innanzi don Rodrigo, in compagnia d'un altro signore; che il primo aveva cercato di provocarlo con canzonature, [...] ma esso, senza dargli retta, aveva affrettato il passo, e raggiunto i compagni. [...] Il giorno dopo, [...] Lucio era nel mezzo dei compagni, con lo sguardo fisso a terra; e l'altro signore sghignazzava, e don Rodrigo diceva: vedremo, vedremo. [...] Al padre Cristoforo, in confessione, madre, — rispose Lucio, con un tono sommesso ma fermo, quasi a chieder scusa."
Copilot	"[...] mentre tornava dalla filanda, ed era rimasto indietro dai suoi compagni, gli era passato innanzi don Rodrigo, in compagnia d'un altro signore; che il primo aveva cercato di trattenerlo con chiacchiere, [...] ma egli, senza dargli retta, aveva affrettato il passo, e raggiunti i compagni. [...] Il giorno dopo, [...] Lucio era nel mezzo dei compagni, con lo sguardo fisso e severo; e l'altro signore sghignazzava, e don Rodrigo diceva: vedremo, vedremo. [...] Al padre Cristoforo, in confessione, padre, — rispose Lucio, con un accento grave di scusa."

This passage shows that the most significant transformation concerns not the characters' language or body movements but the meaning of the event itself. In the original text, the threat faced by Lucia centres on sexual harassment and concerns over honour. In the male versions, however, the same event is largely reinterpreted in terms of social status, power relations, and class superiority. Although the narrative remains the same, the central conflict changes.

This shift is also evident in the expression "*trattenerla con chiacchiere*". ChatGPT reconstructs it in terms of rudeness and disrespect, whereas Gemini emphasises mockery and provocation. In both models, the verbal harassment directed at the female character becomes an expression of power and humiliation when the character is male. Copilot, by contrast, preserves the expression almost unchanged and remains closest to the original. Once again, the models differ in the extent of their intervention.

A similar pattern appears in the expression "*con gli occhi bassi*" ("with her eyes lowered"). ChatGPT interprets the lowered gaze as an attempt to avoid eye contact with Don Rodrigo. Gemini preserves the gesture but presents it as an expression of restrained anger and wounded pride rather than fear and embarrassment. Copilot changes the direction of the gaze entirely, portraying the character as looking directly and sternly at his opponent. Although the three models adopt different solutions, they show the same tendency: body language associated with fear and shame in the female character is reconstructed to express determination, resistance, or anger in the male character.

The parental figure undergoes a similar transformation. In the original text, Lucia refers to her mother; in the male versions, this reference is changed to the father. Thus, the emotional and familial point of reference is reconstructed through the parent considered appropriate to the character's gender.

The expression "*accento soave di scusa*" ("a gentle tone of apology") follows the same pattern. ChatGPT largely abandons the apologetic tone, replacing it with a more distant and reluctant expression. Gemini balances gentleness with determination through the phrase "*sommesso ma fermo*" ("quiet but firm"). Copilot simply replaces *soave* ("gentle") with *grave* ("grave"), making the emotional tone more solemn. Although the degree of intervention differs, all three models move away from a gentle and vulnerable tone of apology towards a more restrained and composed expression.

Example 9

Alessandro Manzoni, I promessi sposi, pp. 257-258.	
Original Passage	«S'alzò, e si mise in ginocchio, e tenendo giunte al petto le mani, dalle quali pendeva la corona, alzò il viso e le pupille al cielo, e disse: — o Vergine santissima! Voi, a cui mi sono raccomandata tante volte, [...] aiutatemi! fatemi uscire da questo pericolo, fatemi tornar salva con mia madre, [...] e fo voto a voi di rimaner vergine; rinunzio per sempre a quel mio poveretto, per non esser mai d'altri che vostra.»
ChatGPT	S'alzò, si mise in ginocchio e, stringendo tra le mani la corona del rosario, levò gli occhi al cielo e disse: — O Vergine santissima! Voi, a cui mi sono rivolto tante volte, aiutatemi! Fatemi uscire da questo pericolo e ricondurmi salvo da mia madre. Se mi concederete questa grazia, consacrerò la mia vita al vostro servizio; rinunzierò al matrimonio e ai desideri del mondo, perché non voglio appartenere ad altri che a Dio.
Gemini	«S'alzò, e si mise in ginocchio, e tenendo giunte al petto le mani, dalle quali pendeva la corona, alzò il viso e le pupille al cielo, e disse: — O Vergine santissima! Voi, a cui mi sono raccomandato tante volte, [...] aiutatemi! fatemi uscire da questo pericolo, fatemi tornar salvo con mia madre, [...] e fo voto a voi di rinunziare per sempre al mondo e a ogni terrena unione; rinunzio per sempre alla mia poveretta, per consacrarmi interamente a voi e non esser mai d'altri che vostro.»
Copilot	«S'alzò, e si mise in ginocchio, e tenendo giunte al petto le mani, dalle quali pendeva la corona, alzò il viso e le pupille al cielo, e disse: — o Vergine santissima! Voi, a cui mi sono raccomandato tante volte, [...] aiutatemi! fatemi uscire da questo pericolo, fatemi tornar salvo con mio padre, [...] e fo voto a voi di rimaner casto; rinunzio per sempre a quella mia poveretta, per non esser mai d'altri che vostro.»

Unlike the previous examples, this passage shows that the models do not introduce gender-based transformations in every part of the text. Ritual elements such as kneeling, joining the hands, holding a rosary, and looking towards heaven are largely preserved across all three models. This suggests that religious rituals are treated as gender-neutral. Rather than rewriting the whole passage, the models intervene only in elements they consider gender-specific.

The most significant transformation concerns the expression "*rimaner vergine*" ("to remain a virgin"). ChatGPT rewrites this passage completely, expressing the vow of virginity as a commitment to a life devoted to God and the renunciation of marriage. Gemini similarly transforms virginity into the renunciation of the world and worldly relationships. Copilot, by contrast, simply replaces *vergine* ("virgin") with *casto* ("chaste"), leaving the rest of the passage largely unchanged. Thus, the three models address the same issue at different levels of intervention. ChatGPT and Gemini redefine the concept, whereas Copilot makes only a minimal change.

ChatGPT also differs from the other models in the figure to whom the vow is directed. In the original passage, the prayer is addressed to the Virgin Mary, whereas ChatGPT redirects it to God. Gemini and Copilot preserve the dedication to Mary. Thus, not only the concept of virginity but also the object of religious devotion changes.

The maternal figure follows a different pattern. The expression "*Fatemi tornar salva con mia madre*" ("Let me return safely to my mother") is preserved by ChatGPT and Gemini, while only Copilot replaces the mother with the father. This shows that the maternal figure is not systematically transformed into the paternal figure. In the previous passages, the transformation was more evident when the mother appeared as a caregiver or the person preparing the character. Here, however, she is largely retained as a figure of security and refuge. The transformation therefore depends not only on the character's gender but also on the mother's narrative function.

A similar difference appears in the expression "*quel mio poveretto*". Gemini and Copilot preserve this affectionate form of address, changing only its gender. ChatGPT removes it altogether and replaces it with the idea of renouncing marriage. Thus, an expression of personal and emotional attachment becomes a more abstract expression of religious devotion.

Conclusion

This study has established a direct connection between feminist theory and the algorithmic outputs of artificial intelligence, showing how patriarchy in the digital age is concealed behind the illusion of technological neutrality. bell hooks' conception of feminism as an intersectional practice of

liberation, Sara Ahmed's feminist *killjoy*, Simone de Beauvoir's theory of the social construction of womanhood, and Virginia Woolf's concept of spatial autonomy provide the analytical framework for examining how large language models (LLMs) process female characters in Italian literature. The masculine bias embedded in the historical and literary texts that constitute the training corpus has led artificial intelligence systems to encode the male subject as the default representation of the human.

The qualitative content and discourse analysis of nine passages from the Italian literary canon shows that ChatGPT, Gemini, and Copilot do not perform gender transformation only at the grammatical level. They also reconstruct characters' attributes and the narrative framework according to patriarchal norms. Where the narrative allows, domestic production and traditionally female forms of labour are replaced by masculinised work such as seafaring, agriculture, and commerce. Likewise, fertility and motherhood become family provision, breadwinning, household leadership, and economic responsibility in the male character. Thus, biological reproduction gives way to public and economic responsibility, reproducing in the digital sphere the spatial division identified by Virginia Woolf.

Consistent with Simone de Beauvoir's argument, qualities such as the desire to be admired, elegance, gentleness, and emotional vulnerability are transformed into respectability, personal grooming, composure, charisma, and authority when assigned to male characters. Expressions of humility, limited knowledge, and embarrassment are reformulated as confidence, practical intelligence, composure, or inexperience. Likewise, visible emotional reactions such as blushing, becoming flustered, or crying are concealed or replaced by anger, seriousness, self-control, and firmness. Male vulnerability is systematically obscured, while sadness and gentleness give way to quiet confidence, dignity, and a sense of duty.

The same pattern appears in dialogue, perceptions of threat, and religious devotion. Sexual harassment, intimidation, and concerns over honour become struggles over status, class, and power. Situations representing existential fear for female characters become questions of prestige and authority for male characters. Religious devotion is also reinterpreted: while ritual forms remain largely unchanged, their meaning shifts towards devotion to God, principled commitment, rational observation, and the renunciation of

worldly desires. Virginity and chastity likewise become abstract expressions of moral resolve.

These findings support Wajcman and Young's intersectional technofeminist argument that artificial intelligence is not a neutral technology but reproduces digital patriarchy through historically biased data and a masculine design culture. Across psychological novels, *verismo*, and historical novels, femininity is associated with passivity, domesticity, emotional dependence, vulnerability, and bodily constraint, whereas masculinity is associated with agency, rationality, public authority, anger, emotional restraint, and economic responsibility. The study therefore demonstrates that feminist intervention is essential for exposing algorithmic systems, questioning their underlying data structures, and contributing to the development of ethical artificial intelligence.

References

Ahmed, S. (2020). *Feminist bir yaşam sürmek* (B. S. Aydaş, Trans.). Sel Yayıncılık.

Enginyurt, D. (2018). Sosyal hareketler sosyolojisi açısından feminizm. *USE Uluslararası Sosyoloji ve Ekonomi Dergisi*, *1*, 36–59. https://izlik.org/JA49BA53HX

Eroğul, N. (2025). ChatGPT'nin toplumsal cinsiyet algısına yönelik söylem analizi. *Nevşehir Hacı Bektaş Veli Üniversitesi SBE Dergisi, YZ (AI) ve Gelecek Perspektifler*, 23-44. https://doi.org/10.30783/nevsosbilen.1742358

Gedik, E. (2020) Dünyada ve Türkiye'de dijital feminizm incelemesi: Gençlerin dijital aktivizm deneyimleri. *Toplum ve Kültür Araştırmaları Dergisi*, 5, 123-136. https://izlik.org/JA89JA93FE

Hooks, b. (2020). *Duygu yoldaşlığı: Kadınların sevgi arayışı* (Ö. Karakaş, Trans.). Bgst Yayınları.

Manzoni, A. (2020). *I promessi sposi*. Feltrinelli.

O'Connor, S., & Liu, H. (2024). Gender bias perpetuation and mitigation in AI technologies: Challenges and opportunities. AI & SOCIETY, 39(5), 2045–2057.

Svevo, I. (1930). *La coscienza di Zeno*. Morreale.

Ün, B. (2023). Virginia Woolf'un "kendine ait bir oda" isimli eseri bağlamında feminist bir manifesto olarak bellek. *Marmara Üniversitesi Kadın ve Toplumsal Cinsiyet Araştırmaları Dergisi*, 7(1), 44-53. https://izlik.org/JA45NG23JY

Verga, G. (2017). *I Malavoglia: Edizione integrale*. Demetra.

Wajcman, J., & Young, E. (2023). Feminism confronts AI: The gender relations of digitalisation. In J. Browne, S. Cave, E. Drage, & K. McInerney (Eds.), *Feminist AI* (pp. 47–64). Oxford University Press.

Ethical Use of Gen-AI in Language Teacher Education

Nazlı Ceren Cirit-Işıklıgil[1]

The Rise of Gen-AI in Language Education

Following its release on November 30, 2022, ChatGPT, a generative artificial intelligence (Gen-AI) chatbot built on a large language model (LLM), reached 1 million active users within 5 days (Dwivedi et al., 2023). In response to the public's rapid uptake, several other Gen-AI chatbots, such as Google Gemini, Microsoft Copilot, Anthropic Claude, and DeepSeek, emerged and have been used in language education since then (Thorne, 2024). This rapid, worldwide adoption has likewise begun reshaping education through human-like dialogue and practical outputs. This technology has specifically made its presence and potential increasingly evident in language classrooms by personalizing learning and automating tasks for both teachers and learners. Teachers can now utilize Gen-AI for lesson planning (Dornburg & Davin, 2025), instructional material design (Kohnke & Zou, 2026), automated writing scoring (Bui & Barrot, 2025; Kim, 2025; Pack et al., 2024), and feedback generation (Guo & Wang, 2024). Students can obtain grammar support (Yang & Chen, 2026), vocabulary support (Abdelhalim & Alsehibany, 2025), speaking practice (Tai & Chen, 2024; Zou et al., 2025), and listening practice (Zhou et al., 2025). The language education field has an intrinsic relationship with Gen-AI, with an important distinction from other disciplines: the language generated by Gen-AI serves as both the medium and the subject matter of the discipline, creating a direct

[1] Hasan Âli Yücel Faculty of Education, Istanbul University Cerrahpaşa, Hasan Âli Yücel Faculty of Education, Department of English Language Teaching, ORCID: 0000-0001-7422-2026, *ceren.cirit@iuc.edu.tr*, Türkiye

link between the two. Thus, this connection entails field-specific risks, necessitating tailored guidance for language teachers.

The Need for Ethical Guidance in Language Education

The widespread adoption of Gen-AI by language teachers for professional purposes has placed a responsibility on teacher educators to equip pre-service teachers (PSTs) and in-service teachers (ISTs) to use this technology critically, an expectation presupposing teacher educators' critical competence. For teacher educators to provide such guidance, the risks of Gen-AI use in language teaching need to be critically evaluated and unpacked. Equipping teachers with the skills to navigate Gen-AI entails far more than having relevant technical skills; it requires aligning with specific pedagogical objectives and critically evaluating the consequences for linguistic inclusivity, student privacy, academic integrity, assessment, and learner agency.

Recently, general educational guidance (European Commission, 2026; Miao & Cukurova, 2024; OECD, 2024a) has proposed fundamental principles for ethical and responsible AI use, and specific guidance for LTE (Kostoulas et al., 2026; British Council, 2026; TESOL International Association, 2026) has also started to emerge. Existing guidance includes general conceptual discussions of equity, privacy, algorithmic bias, user agency, ethical competence, and transparency, but remains mostly fragmented. Thus, further development in language teacher education is needed on ethical issues grounded in discipline-specific assumptions and pedagogical goals, accompanied by contextual interpretations. Therefore, the following section discusses several ethical concerns that need context-specific evaluation.

Linguistic bias and normativity: Despite their variations in theoretical basis, widely accepted frameworks in applied linguistics such as translingualism (Canagarajah, 2013), World Englishes (Kachru, 1985), and English as a lingua franca (ELF) (Jenkins, 2015; Seidlhofer, 2011) likewise argued against the imposition of a "standard" variety of English, such as British or American English, on language learners and rejected the standardized, universal native speaker norms of English. Crystal (2003) highlighted that Kachru's (1985) expanding (foreign language speakers of English) and outer

circle (second language speakers of English) immensely outnumber those of the inner circle (native speakers of English). Despite this evidence, large language models (LLMs) such as ChatGPT cannot be trusted to ensure linguistic inclusivity and are inclined to generate standardized, inner-circle varieties of English. Empirical evidence suggests that users of non-standardized dialects receive biased and even unreliable outputs (British Council, 2026; Fleisig et al., 2024; Hofmann et al., 2024) or that learners of English who show traces of their native languages often receive limited outputs or feedback (Liang et al., 2023; Pack et al., 2024; Yang et al., 2025). This issue is a critical ethical concern in language teacher education that requires specific guidance for stakeholders in the field of language education, since PSTs are not taught to treat interlanguage (Selinker, 1972), a learner language, as something to be corrected or penalized but as a systematically evolving linguistic system (Tarone, 2018). Thus, ethical guidance for teachers to evaluate Gen-AI-generated feedback on learners' spoken or written practice through a critical lens is needed so that learners' naturally developing target-language output and its sociolinguistic variations are not mistaken for genuine errors that cause communication breakdowns.

Student privacy: Teachers frequently ask language learners to complete personalized tasks and submit assignments (e.g., describing themselves and their families and discussing their daily lives) (Harmer, 2015), which can elicit personally identifiable information. This opens a debate about how all this sensitive information is handled when uploaded to Gen-AI. Is it left to a teacher's discretion, or are privacy issues governed by clear protocols? Although the British Council (2026) advises teachers to anonymize student data before uploading their assignments to AI, Gen-AI language assessment procedures should be regulated at the governmental and institutional levels (Pack et al., 2024).

Academic integrity and learner authorship: The European Commission (2026, p. 21) sees academic integrity as one of the core components of the learning process and describes it as "using AI in an honest and ethical manner, such as not misrepresenting contributions, giving proper attribution to the ideas and work of others, and designing assessments that remain valid in the era of generative AI." In line with the European Commission's perspective, academic integrity is one of the most frequently cited concerns in

educational settings (Yan et al., 2024). Now, learners are bridging the gap between what they are able to produce in L1 and what they can currently produce in L2 thanks to AI feedback, translation, and paraphrasing tools. In response, plagiarism detection tools (Turnitin, iThenticate) are now extended to detect AI-generated content, and other tools (GPTZero) have been developed specifically for this purpose. These detection tools, however, systematically misclassify non-native English texts as "AI-generated" (Liang et al., 2023). Then, in case of standard AI detection, language learners may be falsely accused. Kohnke et al. (2023, p. 544) likened the competition between AI's capabilities and detection efforts to a "cat and mouse" situation. The TESOL International Association's (2026) guidelines emphasize the necessity of explicit language education policies to inform such decisions. Therefore, questions arise in relation to authorship, attribution, and originality: To what extent can student text be treated as original? How should AI assistance be acknowledged? The boundaries between learner authorship and Gen-AI-assisted authorship can easily blur. Using field-specific principles, answers to these questions should be carefully crafted with regard to students' stages of linguistic development and the requirements of a task.

Assessment validity: The European Commission's 2026 guidelines identify assessment as one of the key considerations for the ethical implications of AI use. For a discipline like language education whose regular practices involve written and spoken production, ethical issues regarding assessment become critical. Rather than banning AI use or trying to detect AI assistance, redesigned assessment procedures should be envisioned to make the writing process (i.e., drafting, revising, reflecting) transparent and thereby serve as a learning opportunity (Elturki, 2026). The British Council (2026) similarly suggests that transparency could be beneficial, as learners can also learn from others' prompts. For that purpose, an alternative process-oriented assessment can be adopted for written pieces by enabling the visible documentation of instructor feedback and students' conversations with AI (Elturki, 2026).

Learner agency and critical Gen-AI engagement: PSTs need training on how to operate Gen-AI critically, question its outputs, and recognize that it can produce biased and inherently limited responses. Therefore, this

competence can be transferred to their students for responsible adoption. This would turn Gen-AI into an asset for language teaching and learning. Following a course intervention, Moorhouse et al. (2024) reported increased awareness among pre-service teachers regarding AI use, viewing it as a supporting tool rather than an infallible authority. For learners, rather than cognitive overreliance on what AI offers unquestioningly, retaining agency over their language development and investing real effort in learning are essential to achieving proficiency in the target language.

Scope and Aim of the Chapter

This chapter aims (1) to discuss ethical concerns related to the use of Gen-AI in language education and (2) to propose context-sensitive principles that can guide language teacher educators to be role models and mentor PSTs and ISTs' paths. The chapter focuses specifically on Gen-AI because what this technology can do—generating, translating, transforming, assessing, and responding to language—overlaps quite directly with the object and medium of language education. In light of these two main aims, the chapter proposes principles that are derived from the evidential basis and a conceptual synthesis of existing frameworks and guidelines, in general (e.g., European Commission, 2026; Miao & Cukurova, 2024) and specifically in language education (e.g., British Council, 2026; TESOL International Association, 2026). Gen-AI literacy frameworks for language teachers have recently appeared in the literature, and the principles proposed here are consistent with the competencies those frameworks specify. This chapter focuses on the competencies these frameworks measure but fail to explicate in relation to the pedagogical Gen-AI decisions language teachers need to make in a specific context. The chapter, therefore, is not intended as a proposal for an established competency framework but rather presents disciplinary insights to facilitate pedagogically sound ethical decision-making and to guide future research on empirical validation.

Key Terms and Definitions

Below, the definitions of the key terms around which the chapter revolves are given, along with their widely accepted definitions that align with the scope of this chapter:

Artificial intelligence (AI): AI as an umbrella term is defined as a machine-based system that has explicit or implicit objectives, can make inferences based on the input given, and is capable of producing outputs in the form of predictions, content, recommendations, or decisions that may have an impact on real or virtual contexts (OECD, 2024b). This chapter intentionally uses the OECD definition because it is comprehensive enough to encompass earlier AI applications (i.e., rule-based grammar checkers and intelligent tutoring systems), which have long been used in language education and are now used by learning-based systems.

Generative artificial intelligence (Gen-AI): As an AI technology, Gen-AI produces new automated content by drawing on existing content in response to direct user-generated prompts (Miao & Cukurova, 2024). Here, Gen-AI refers to systems, most often built on LLMs, that generate multimodal text in response to everyday dialogic prompts. This chapter focuses specifically on Gen-AI, a subset of AI, because this technology is within reach of teachers and students, and it not only classifies or retrieves language but also generates it. Also, Gen-AI produces fast, fluent, and open-ended language from evolving, comprehensive corpora instead of a stable, compiled database. Opportunities to personalize learning, design language teaching materials, produce lesson content, and increase student interaction and engagement make Gen-AI distinctive and applicable in the language education context. Along with these opportunities, the use of Gen-AI in language teaching and learning has ethical consequences, as discussed in this chapter.

AI literacy: In broad terms, AI literacy refers to comprehension and awareness of AI's affordances and what is required to participate in wider AI conversations, along with making sound choices about use and implications (Southworth et al., 2023). Ng et al. (2021) categorized the construct into four main aspects: (1) knowing and understanding AI, (2) using and applying it, (3) evaluating and creating with it, and (4) addressing its ethical issues. The construct has since been reconceptualized following the

introduction of Gen-AI tools. For instance, Annapureddy et al. (2025) proposed twelve Gen-AI literacy competencies that range from conceptual knowledge to prompting, evaluating, and critically engaging with ethical concerns. In language education, Kohnke et al. (2023) identified specific digital competencies that learners and teachers need to use ChatGPT ethically and effectively. Specifically for language teachers, Huang et al. (2025) operationalized Gen-AI literacy across five sub-dimensions: technical proficiency, critical evaluation, communication proficiency, creative application, and ethical competence. They tested the construct in a twelve-week teacher education course and provided insights for language teacher education programs. Based on AI literacy competency studies, this chapter particularly characterizes AI literacy as the capability to understand the technical, pedagogical, and critical affordances of Gen-AI in a given language and to use it responsibly and purposefully, guided by objectives and sociolinguistic criteria.

AI Ethics: The field of AI ethics concerns the moral principles and practices that should guide the design, use, and deployment of AI technologies. Recent literature (Nguyen et al., 2023; Yan et al., 2024) reveals recurring principles for AI ethics: privacy, transparency, accountability, inclusiveness, and human agency. General guidelines from UNESCO and the European Commission translate these principles into practice for teachers (European Commission, 2026; Miao & Holmes, 2023).

Ethical Competence: In this chapter, it refers to a teacher's ability to critically and professionally evaluate the ethical aspects of Gen-AI-related decisions in educational practice and to take contextually appropriate actions grounded in discipline-specific expertise. In UNESCO's competency framework for teachers, Miao and Cukurova (2024, p. 23) propose five core aspects of essential competencies, two of which are "human-centered mindset" and "ethics of AI." Teachers need to develop agency in fostering social responsibility and attitudes, as well as participate in the practical application of ethical rules. This chapter treats ethical competence as a disposition rather than a mere list of tasks to achieve. When a language teacher can critically evaluate and determine the appropriateness of Gen-AI feedback for a specific context, for instance, penalizing students for using an outer-circle spelling, that teacher demonstrates ethical competence.

Theoretical Foundations

The frameworks, guidelines, and empirical studies reported in this section were identified through searches in databases such as Scopus, Web of Science, and Google Scholar, covering 2020-2026. The synthesis was purposive, and five aforementioned ethical concerns emerged from recurring themes across these documents. The synthesis included documents that explicitly laid out ethical principles or competencies for teachers' or learners' AI use, as well as empirical studies that measured, modelled, or taught AI or Gen-AI literacy with an explicit ethical component in a language education context.

General Education Guidelines and Policy Frameworks

Among global policy frameworks, UNESCO's *Guidance for generative AI in education and research* (Miao & Holmes, 2023) is the first global guidance from UNESCO that specifically handles Gen-AI in education. Based on a human-centered approach, this document addresses national authorities, policymakers of teacher education programs, teachers, and researchers on Gen-AI use and raises awareness of data protection, student privacy, and the critical evaluation of Gen-AI affordances.

UNESCO also proposed an *AI competency framework* for teachers (AI CFT) and students in 2024. The overall aim of these frameworks is to frame AI use not merely as technical proficiency but as a critical, ethical, and responsible practice. The teacher framework systematically categorizes teacher AI competencies, including ethical competence as a core component (Miao & Cukurova, 2024). In this framework, teachers are expected to actively and critically check Gen-AI outputs, serve as designers and facilitators in AI-supported learning, protect an ethical, safe practice, and model lifelong learning about AI. The framework acknowledges the need to train teachers to meet these expectations and take responsibility by understanding how to minimize Gen-AI risks and make full use of its benefits. For that, this document provides a baseline for teacher educators to design and plan AI training courses.

Another important addition for this group of documents would be *Empowering Learners for the age of AI* (AILit) Framework (OECD / European Commission, 2026). This framework is particularly relevant because it

complements UNESCO's AI teacher competency framework by guiding learners to make autonomous decisions and think critically about appropriate AI output. In this framework, AI ethics is not an isolated checklist item but woven into the pedagogical process itself. The learners' competencies are organized under process-relevant headings such as engaging with, creating with, managing, and shaping AI, where ethical considerations are spread all over. Last but not least, it further guides the language education field by addressing issues such as originality, authorship, privacy, bias, and human agency.

The European Commission's revised *Guidelines on the ethical use of artificial intelligence and data in teaching and learning for educators* (2026) are distinctive because they have more explicit references to language-related concepts (i.e., multilingualism and diversity) than earlier general education frameworks. These guidelines make academic integrity and assessment validity integral parts of the framework along with context-based scenarios and guiding questions for ethical decisions. However, like other frameworks and guidelines, they also address education in general and therefore do not provide ethical baselines for cases where teachers should handle AI treatment of learners' interlanguage, errors, or biases against nonstandard varieties of English. This situation is where a language-specific framework becomes essential.

Most recently, the Council of Europe (2026) published a recommendation on artificial intelligence literacy in which they promote critical AI engagement and identify AI use as a much deeper capacity than technical proficiency, which is quite parallel with the focus of this chapter.

Overall, these frameworks prioritize the aforementioned issues associated with ethical AI use; however, because they are designed for educational disciplines in general, teachers must interpret how they apply to a specific discipline. In language education, when making decisions and judgments about these concerns, experience and knowledge across multiple constructs (i.e., linguistics, sociolinguistics, acquisition, and assessment) come into play.

Language Education Field-Specific Frameworks

Well-known models such as TPACK and SAMR do not include the ethical dimension of AI, even though Kohnke and Zou (2025) used these frameworks to critically integrate AI into language teacher education. Even though, they were not from the language education field, Çelik (2023) proposed Intelligent-TPACK which comprises an ethical knowledge component, and Falloon (2020) proposed a teacher digital competency framework that treats ethical knowledge as a separate dimension. Therefore, even though language-specific policy frameworks and guidelines are scarce, there are recent updates from three main professional organizations that aimed to fill this gap: (1) Kostoulas et al. (2026) with the Council of Europe's AI Lang project, (2) the British Council (2026) guidelines, and (3) TESOL International Association (2026) guidelines.

The Council of Europe's European Centre for Modern Languages (ECML), its specialist body for language education, proposed the *Framework for the use of AI in language education* (Kostoulas et al., 2026) for the AI for Language Education project, which runs from 2024 to 2027. This framework, one of the project's key outputs so far, proposes four principles (i.e., Safe, Responsible, Purposeful, and Human-centred) for language teachers and teacher educators. The principles are categorized through eight guidelines and a set of competence descriptors. This framework closely relates to the discussions in this chapter; therefore, it is the most relevant resource to date. Given that it discusses AI use beyond technical competence and situates it within the pedagogical realities of language education, the framework relates to the chapter's main arguments in various aspects since it (1) underlines learner data protection (safe), (2) draws teachers' attention towards ethical and societal concerns such as equity and environmental sustainability (responsible), (3) foregrounds the necessary linking between the technology and pedagogical aims (purposeful), and (4) leads teachers and teacher educators to approach AI from a critical perspective and be open-minded for continuous evaluation of the practices (human-centred). Therefore, these principles align with the current chapter's argument that Gen-AI competence requires teachers to follow professional judgement rather than following context-independent fixed rules. The current chapter, however, makes contributions to the literature since it addresses Gen-AI

specifically and interprets ethical issues via disciplinary constructs grounded in applied linguistics and second language acquisition (SLA). For instance, this chapter asks teacher educators and teachers what responsible feedback is when Gen-AI encounters interlanguage errors. In addition, this chapter's intended audience is teacher educators. Therefore, rather than mainly proposing what teachers should know and do, the principles propose how teacher educators can help PSTs and ISTs to develop a specific competence, in part by making their own Gen-AI decision-making visible in their courses.

The British Council (2026) offers practical guidance on data anonymization, evaluating Gen-AI feedback through a critical lens, and enabling prompt visibility to learners. TESOL International Association (2026) emphasizes the necessity of specific AI policies in language education and for raising teacher awareness of the affordances and limitations of the AI tools they use.

Alongside these guidance documents from well-known organizations, empirical studies on competency models have also emerged in recent research. In an intervention study, Huang et al. (2025) examined Gen-AI literacy across five dimensions, including ethical competence, among 90 PSTs of English. The experimental group that completed a twelve-week Gen-AI literacy course had significant gains in all Gen-AI literacy dimensions. Pan and Wang (2025) constructed and validated a scale for AI literacy for teachers in the English as a Foreign Language (EFL) context, which includes ethics as a subscale. Other studies (Ou et al., 2024; Wang & Wang, 2025) focused on developing critical AI literacy models for L2 writers. Moorhouse et al. (2024) found that an eleven-week course on Gen-AI competence helped PSTs to improve pedagogical competence and critical awareness; however, it did not enhance PSTs' capacity to guide their students' ethical and responsible AI use. The study shows that ethics is measured as a teacher's AI literacy component; however, it is not linked to what teachers do when an ethical challenge arises in their classrooms.

Proposed Principles

Based on the ethical concerns identified in the relevant literature and frameworks, five main principles are proposed to guide language teacher educators in preparing PSTs and ISTs for the responsible Gen-AI use in language classrooms. The principles are not hierarchical; rather, they are intended to operate in combination. When an ethical issue arises, different principles may matter more depending on the issue and context.

Principle 1: Linguistic Fairness

Teacher educators should raise awareness that Gen-AI-generated outputs or feedback may include linguistic bias; therefore, they should prepare PSTs and ISTs to approach the output cautiously and critically, rather than accepting it without question. This principle is grounded in various foundational research and empirical studies: evidence of dialect discrimination in LLMs (Fleisig et al., 2024; Hofmann et al., 2024), native language influence that puts learners at a disadvantage (Liang et al., 2023), the ongoing refusal of native speaker norms (Canagarajah, 2013; Crystal, 2003; Jenkins, 2015; Kachru, 1985), and the stance towards interlanguage as a developing system rather than a deficit (Selinker, 1972; Tarone, 2018).

To operationalize this principle, PSTs and ISTs first assess anonymized samples of student writing without any AI assistance and then use a Gen-AI chatbot to assess the same texts. They subsequently compare the chatbot's feedback with their personal assessment. Teacher educators facilitate this comparison session through critical guiding questions to identify which Gen-AI-generated feedback, error correction (i.e., are they errors or variations?), or vocabulary suggestions (i.e., how does it treat particular vocabulary choices for any register, culture, or non-native influence?) they accept or reject, and discuss the rationale behind these choices. Then, teacher educators can moderate a follow-up discussion on the chatbot's treatment of language varieties and culturally situated language use to check whether Gen-AI favors some varieties or cultural references over others.

Principle 2: Learner Data Protection

Teacher educators should teach PSTs and ISTs to treat learner data as personal data and protect it according to its sensitivity. In line with existing guidance, this principle requires PSTs and ISTs to remove identifying information from learners' work before seeking AI assistance, as the British Council (2026) suggests. In language education, this issue is particularly salient because teachers often elicit personal information through writing and speaking tasks and assignments that may include learners' names, photographs, contact details, descriptions of family members, places of residence, and habits (Harmer, 2015). Although teacher educators play a significant role in preparing PSTs and ISTs to identify and train on this matter, the responsibility should not be left to teachers' discretion alone; governmental- and institutional-level regulations must be established.

In practice, teacher educators can walk PSTs and ISTs through coursebook activities to identify tasks that pose minimal risks and those that may elicit personal, sensitive information. They could also facilitate discussions about different levels of sensitivity and extend practice with fabricated sample learner texts that include various types of personal information to help them identify what information is sensitive (i.e., frequently visited locations) and what carries little risk (i.e., food preferences). These sorts of activities can give pre-service teachers opportunities to develop context-sensitive judgments about learner data and decide whether additional protective precautions are necessary.

Principle 3: Developmental Authorship

Following the European Commission's (2026) definition of academic integrity, teacher educators should guide PSTs and ISTs to define learner authorship by considering learners' language development stage and task requirements. In addition, PSTs and ISTs need to clearly state how much AI involvement is tolerable at the start of the task, since prohibiting learners' AI use cannot actually prevent it (Kohnke et al., 2023). Because a task a learner completes must reflect their own learning, an assessment instrument is needed to define "AI involvement" and "reflecting own learning," and the AI Assessment Scale developed by Perkins et al. (2024) can serve this purpose. Teacher educators can guide teachers to use this practical

instrument, which equips them to identify the permitted levels of AI involvement per task according to the learning outcomes. For further practice, teacher educators can ask teachers to design a task and justify its objectives and the level of AI involvement it permits, using a written AI-use statement.

Principle 4: Process-visible Assessment

Teacher educators should promote visibility of the assessment process by modeling its design. PSTs and ISTs should be trained in a way that their assigned tasks allow the transparency of the learners' interaction with a Gen-AI chatbot rather than attempts at AI-assistance detection on the task product. This principle follows directly from Elturki's (2026) process-based AI-involved writing pedagogy, the British Council's (2026) proposal on shared prompts among learners to enable learning from one another, and the European Commission's (2026) assessment validity argument, which states that assessment remains valid as long as it assesses the intended competence, even when Gen-AI is involved. This argument is particularly important in language education, since an AI-polished text may not accurately reflect learners' linguistic competence when that is the target construct being assessed. By contrast, if the task intends to assess revision ability, teachers could accept Gen-AI feedback as long as they can see learners' decision-making process regarding what they accept or reject. Therefore, teacher educators should model such visible reasoning for PSTs and ISTs so AI use does not threaten assessment validity but instead supports it.

Principle 5: Reflective Agency

Teacher educators should cultivate critical Gen-AI engagement as a routine professional practice, not just among PSTs and ISTs but also in their own work. Following from Sperling et al.'s (2024) call to restore professional judgment to the centre of AI literacy in teacher education and from Holmes et al.'s (2022) distinction between doing ethical things and doing things ethically, reflective agency is one's capacity to examine their own use of Gen-AI, question and evaluate its outputs, and make informed decisions on its use. This principle is distinctive because it allows the other four principles to apply across different pedagogical and ethical situations. It is particularly relevant to teacher educators since they also frequently feel

unconfident with Gen-AI (Moorhouse & Kohnke, 2024). Teacher educators can actually put this principle into practice by making their own Gen-AI use visible to PSTs and ISTs and by explaining the rationale behind evaluating, questioning, revising, accepting, or rejecting AI-generated outputs.

Implications

For language teacher education, one notable implication is that AI ethics cannot be treated as an isolated component; rather, the proposed context-specific principles need to be embedded across teacher educators' own practices. Second, it is imperative to provide teacher educators the support they need for ethical Gen-AI use. Teacher educators need to be provided with adequate support to feel confident in using, evaluating, and teaching with Gen-AI if they are to develop PSTs' and ISTs' ethical Gen-AI use. Third, the principles need to be localized. Although the principles were intended to be broad enough to apply across contexts, the chapter's evidential basis mostly comes from Hong Kong, Europe, or North America. In other contexts, where English is not the target language and where Gen-AI does not produce reliable outputs in the local or target language, the same ethical principles may lead to different pedagogical decisions. Localizing the principles, therefore, is critical to ensure they address the linguistic, cultural, technological, and institutional realities of other language education contexts.

Future research is needed to validate the proposed principles, as they are grounded in a limited and recently emerging evidential base. Further studies should investigate how Gen-AI systems provide feedback on students' interlanguage at various proficiency levels and across different native-language backgrounds. Finally, longitudinal research should examine whether process-oriented, transparent assessment practices actually provide sustained benefits for the learners.

National and international policies need to recognize that language education presents discipline-specific challenges; therefore, this field needs relevant Gen-AI guidance. Institutions should provide clear data-handling protocols to guide teachers so learner data protection responsibility does not rest solely on individual teachers' shoulders. Furthermore, Gen-AI competence

should be added as an expected outcome in accreditation and curriculum standards for LTE programs and should use discipline-specific treatments and terminology rather than generic descriptors. Further research is needed to examine how Gen-AI performs in languages other than English, particularly for contexts where training data is limited, since narrow training data might increase the risk of inaccurate, biased, or culturally inappropriate outputs. As policies reach classrooms through teacher education, teacher educators are the key audience for translating these principles into discipline-sensitive pedagogical decisions.

Conclusion

Recent research has yielded more comprehensive and relevant ethical guidelines and frameworks for language education than those available before ChatGPT's public launch. Because language is both the medium and the object of the field, Gen-AI use requires specific ethical principles. Based on the discipline-specific concerns (linguistic bias, privacy, academic integrity, assessment validity, and critical AI engagement) elaborated, five main principles were proposed: linguistic fairness, learner data protection, developmental authorship, process-visible assessment, and reflective agency. Despite the scarcity of field-specific frameworks and few recent empirical studies underpinning these proposed principles, they provide a starting point for teacher educators who need guidance for the PSTs and ISTs they train and advise. The chapter's primary contribution is to shift the discussion of ethical Gen-AI use from mere adherence to a fixed set of ethical principles toward discipline-sensitive professional judgment, where ethical decisions are interpreted based on the pedagogical, linguistic, and sociolinguistic realities of language education contexts.**Declaration on the Use of Artificial Intelligence**

The author used ChatGPT and Claude to verify reference metadata, edit language, and refine style.

Conflicts of Interest

The author declares no conflicts of interest.

Data Availability Statement

This is a conceptual/theoretical chapter, and no new datasets were generated.

Ethics Statement

Ethical approval and informed consent were not required, as this conceptual chapter did not involve human participants, animal subjects, or the collection of primary data.

References

Abdelhalim, S. M., & Alsehibany, R. A. (2025). Integrating ChatGPT for vocabulary learning and retention: A classroom-based study of Saudi EFL learners. *Language Learning & Technology, 29*(1), 1–24.

Annapureddy, R., Fornaroli, A., & Gatica-Perez, D. (2025). Generative AI literacy: Twelve defining competencies. *Digital Government: Research and Practice, 6*(1), 1–21.

British Council. (2026). *AI guidelines for teachers*. TeachingEnglish. https://www.teachingenglish.org.uk/sites/teacheng/files/2026-01/TeachingEnglish_AI_guidelines_for_teachers.pdf

Bui, N. M., & Barrot, J. S. (2025). ChatGPT as an automated essay scoring tool in the writing classrooms: How it compares with human scoring. *Education and Information Technologies, 30*, 2041–2058.

Canagarajah, S. (2013). *Translingual practice: Global Englishes and cosmopolitan relations*. Routledge.

Çelik, İ. (2023). Towards Intelligent-TPACK: An empirical study on teachers' professional knowledge to ethically integrate artificial intelligence (AI)-based tools into education. *Computers in Human Behavior, 138*, Article 107468.

Council of Europe. (2026). *Recommendation CM/Rec(2026)12 of the Committee of Ministers to member states on artificial intelligence literacy*. Council of Europe. https://search.coe.int/cm/eng?i=09125948802cde8e

Crystal, D. (2003). *English as a global language* (2nd ed.). Cambridge University Press.

Dornburg, A., & Davin, K. J. (2025). ChatGPT in foreign language lesson plan creation: Trends, variability, and historical biases. *ReCALL, 37*(3), 332–347.

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., ... Wright, R. (2023). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges, and implications of generative conversational AI for research, practice, and policy. *International Journal of Information Management, 71*, Article 102642.

Elturki, E. (2026). AI-aware process writing pedagogy: Rethinking the process-based approach. *TESOL Journal, 17*(2), Article e70127.

European Commission, Directorate-General for Education, Youth, Sport, and Culture. (2026). *Guidelines on the ethical use of artificial intelligence and data in teaching and learning for educators* (Revised ed.). Publications Office of the European Union.

Falloon, G. (2020). From digital literacy to digital competence: The teacher digital competency (TDC) framework. *Educational Technology Research and Development, 68*(5), 2449–2472.

Fleisig, E., Smith, G., Bossi, M., Rustagi, I., Yin, X., & Klein, D. (2024). Linguistic bias in ChatGPT: Language models reinforce dialect discrimination. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 13541–13564). Association for Computational Linguistics. https://aclanthology.org/2024.emnlp-main.750/

Guo, K., & Wang, D. (2024). To resist it or to embrace it? Examining ChatGPT's potential to support teacher feedback in EFL writing. *Education and Information Technologies, 29*(7), 8435–8463.

Harmer, J. (2015). *The practice of English language teaching* (5th ed.). Pearson Education.

Hofmann, V., Kalluri, P. R., Jurafsky, D., & King, S. (2024). AI generates covertly racist decisions about people based on their dialect. *Nature, 633*, 147–154.

Holmes, W., Porayska-Pomsta, K., Holstein, K., Sutherland, E., Baker, T., Buckingham Shum, S., Santos, O. C., Rodrigo, M. T., Cukurova, M., Bittencourt, I. I., & Koedinger, K. R. (2022). Ethics of AI in education: Towards a community-wide framework. *International Journal of Artificial Intelligence in Education, 32*(3), 504–526.

Huang, T., Wu, C., & Wu, M. (2025). Developing pre-service language teachers' GenAI literacy: An interventional study in an English language teacher education course. *Discover Artificial Intelligence, 5*(1), Article 163.

Jenkins, J. (2015). Repositioning English and multilingualism in English as a lingua franca. *Englishes in Practice, 2*(3), 49–85.

Kachru, B. B. (1985). Standards, codification, and sociolinguistic realism: The English language in the Outer Circle. In R. Quirk & H. G. Widdowson (Eds.), *English in the world: Teaching and learning the language and literatures* (pp. 11–30). Cambridge University Press.

Kim, Y. (2025). Automated essay scoring with GPT-4 for a local placement test: Investigating prompting strategies, intra-rater reliability, and alignment with human scores. *TESOL Quarterly, 59*(S1), S318–S329.

Kohnke, L., & Zou, D. (2025). Artificial intelligence integration in TESOL teacher education: Promoting a critical lens guided by TPACK and SAMR. *TESOL Quarterly, 59*, S267–S278.

Kohnke, L., & Zou, D. (2026). Developing GenAI literacy in EAP teacher education: A case study of materials design and professional identity. *Journal of English for Academic Purposes, 83*, Article 101764.

Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for language teaching and learning. *RELC Journal, 54*(2), 537–550.

Kostoulas, A., Dubrac, A.-L., Ljalikova, A., & Pölzleitner, E. (2026). *Framework for the use of AI in language education* [Beta version]. European Centre for Modern Languages of the Council of Europe. https://www.ecml.at/en/ECML-Programme/Programme-2024-2027/AI-for-language-education

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns, 4*(7), Article 100779.

Miao, F., & Cukurova, M. (2024). *AI competency framework for teachers*. UNESCO.

Miao, F., & Holmes, W. (2023). *Guidance for generative AI in education and research*. UNESCO.

Moorhouse, B. L., & Kohnke, L. (2024). The effects of generative AI on initial language teacher education: The perceptions of teacher educators. *System, 122*, Article 103290.

Moorhouse, B. L., Wan, Y., Wu, C., Kohnke, L., Ho, T. Y., & Kwong, T. (2024). Developing language teachers' professional generative AI competence: An intervention study in an initial language teacher education course. *System, 125*, Article 103399.

Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence, 2*, Article 100041.

Nguyen, A., Ngo, H. N., Hong, Y., Dang, B., & Nguyen, B.-P. T. (2023). Ethical principles for artificial intelligence in education. *Education and Information Technologies, 28*(4), 4221–4241.

OECD. (2024a). *Education Policy Outlook 2024: Reshaping teaching into a thriving profession from ABCs to AI*. OECD Publishing.

OECD. (2024b). *Explanatory memorandum on the updated OECD definition of an AI system* (OECD AI Papers No. 8).

OECD / European Commission. (2026). *Empowering learners for the age of AI: An AI literacy framework for primary and secondary education*. OECD Publishing.

Ou, A. W., Khuder, B., Franzetti, S., & Negretti, R. (2024). Conceptualising and cultivating Critical GAI Literacy in doctoral academic writing. *Journal of Second Language Writing, 66*, Article 101156.

Pack, A., Barrett, A., & Escalante, J. (2024). Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability. *Computers and Education: Artificial Intelligence, 6*, Article 100234.

Pan, Z., & Wang, Y. (2025). From technology-challenged teachers to empowered digitalized citizens: Exploring the profiles and antecedents of teacher AI literacy in the Chinese EFL context. *European Journal of Education, 60*(1), Article e70020.

Perkins, M., Furze, L., Roe, J., & MacVaugh, J. (2024). The Artificial Intelligence Assessment Scale (AIAS): A framework for ethical integration of generative AI in educational assessment. *Journal of University Teaching and Learning Practice, 21*(6).

The Matrix of the Invisible: Distant Reading and Artificial Intelligence, and the Combinatorial Architecture of Italo Calvino's *Città invisibili*

Cristiano Bedin[1]

1. Introduction: Theoretical and Methodological Framework

In the eighth chapter of *Invisible Cities* (Le città invisibili, 1972), Italo Calvino stages an emblematic encounter between Marco Polo and the Great Khan Kublai, an exchange that unfolds in complete silence:

> From the foot of the Great Khan's throne a majolica pavement extended. Marco Polo, mute informant, spread out on it the wares he had brought back from his journeys to the ends of the empire: a helmet, a seashell, a coconut, a fan. Arranging the objects in a certain order on the black and white tiles, the ambassador tried to depict for the monarch's eyes the vicissitudes of his travels, the conditions of the empire. Kublai was a keen chess player; following Marco's movements, he observed that certain pieces implied or excluded the vicinity of other pieces. He thought: "If each city is like a game of chess, the day when I have learned the rules, I shall finally possess my empire, even if I shall never succeed in knowing all the cities it contains." [...] Contemplating these essential landscapes, Kublai reflected on the invisible order that sustains cities. At times he thought he was on the verge of discovering a coherent, harmonious system underlying the infinite deformities and discords, but no model could stand up to the comparison with the game of chess. Perhaps [...] it would suffice to play a game according to the rules, and to consider each successive state of the board as one of the countless forms that the system of forms assembles and destroys. (Calvino, 2023: 109–110)

[1] Assoc. Prof. Dr., Istanbul University, Faculty of Letters, Department of Italian Language and Literature, ORCID: 0000-0001-6992-244X, *cristiano.bedin@istanbul.edu.tr*, Türkiye

Faced with an empire whose extent can no longer be fully grasped, the emperor turns to the geometric order of the chessboard. On its black-and-white surface, Kublai seeks to represent the complexity of his cities through a system of rules, treating their relations as combinations generated by the possible moves of the game. The contrast between the disorder of the cities and the relations that the chessboard makes visible extends beyond Kublai's reflections. It points toward a problem that recurs throughout Calvino's writing: how a finite set of rules can generate forms whose variety exceeds any single representation of them.

His engagement with Oulipo and his interest in the formal procedures of the sciences are relevant (Aragona, 2008). These interests help explain why Calvino repeatedly describes literary composition in terms of rules, combinations, and discrete elements, using "discrete" in its mathematical sense to refer to distinct elements (Calvino, 1986: 6).

Traditional close reading, however, poses a different problem when the object of analysis is the structure of the work as a whole. In much of this scholarship, attention tends to focus on the peculiar spatial organization of cities such as Bauci, Valdrada, and Esmeralda (cf. the extensive bibliography on Invisible Cities, including Balice, 1986; Ossola 1988; Ravazzoli 1987; Kuon 1987; Barenghi 1992; Segre 1994; Zancan 1996; Bloom 2000; Battistini 2001; Barenghi 2002; Brera 2011). These studies have examined the aesthetic, symbolic, and combinatorial dimensions of individual cities. This approach, generally associated with close reading, has proven productive for individual passages and textual nuance. Its focus on single passages, however, can make broader patterns within the work more difficult to identify (Janicke et al. 2017: 227). The linear sequence of reading can also make relationships among distant passages and recurring patterns across the work difficult to perceive.

Moretti's concept of distant reading provides a possible response to this problem (2000; 2005; 2013). For Moretti, the study of large-scale literary structures requires moving beyond the individual text and examining patterns that become visible only across a wider corpus (Moretti, 2005: 1). In the case of Invisible Cities, Moretti's approach can offer a way of examining the angles and refractions of the textual crystal constructed by Calvino.

The metaphor of the chessboard also appears in Calvino's essay Cybernetics and Ghosts (1967):

> Just as no chess player will ever live long enough to exhaust all the combinations of possible moves for the thirty-two pieces on the chessboard, so we know (given the fact that our minds are chessboards with hundreds of billions of pieces) that not even in a lifetime lasting as long as the universe would one ever manage to make all possible plays. But we also know that all these are implicit in the overall code of mental plays, according to the rules by which each of us, from one moment to the next, formulates his thoughts, swift or sluggish, cloudy or crystalline as they may be (Calvino, 1986: 6–7).

The passage presents the mind in terms that are close to a cybernetic model: a limited number of rules can produce a potentially unlimited range of configurations. For Calvino, thought can be described in terms of information processing: the complexity of experience thus depends on the relations among discrete elements and on the rules through which they can be combined. A similar logic can be observed in the structure of *Invisible Cities*. Kublai's palace becomes a space in which the information contained in Marco Polo's reports is organized and evaluated, while the cities themselves can be understood as unrealized possibilities through which desires, memories, and fears are projected onto physical space (Lima, 2023: 193-196).

This perspective also makes Moretti's graphs, maps, and trees relevant to the analysis of Calvino's work (Moretti, 2005). The development of AI and deep-learning methods has since expanded the range of computational techniques available to the *Digital Humanities* (Ciotti, 2021: 195–196). When Franco Moretti developed his theory of distant reading in the early 2000s, the computational tools available to literary scholars were still largely based on classical lexicometrics (cf. Moretti, 2000). This first wave of quantitative criticism relied on frequency counts, keyword extraction, and mapping of literal co-occurrences (Moretti, 2000, 2013; Jockers, 2013; Underwood, 2017), effectively reducing texts to a "bag of words," that is, collections of lexical items considered largely independently of their syntactic organization (cf. Ciotti, 2021: 203). Large Language Models (LLMs) and Natural Language Processing (NLP) algorithms now allow scholars to work with textual representations that go beyond simple frequency counts, including vector-based representations of semantic relationships. Although

Ciotti (2021; 2023) has highlighted limitations in the use of AI for literary analysis, these tools remain useful provided their outputs are subject to the scholar's critical evaluation. In this sense, AI might be understood as a computational extension of the critic's gaze, capable of registering both semantic structures and affective dimensions of texts (Romele, Severo, and Furia, 2020; Wellner, 2020).

In recent years, the adoption of neural NLP models and vector-based representations of language has brought greater attention to the semantic and relational dimensions of texts in the Digital Humanities (Wevers & Koolen, 2020; Hu & Amaral, 2022; Leavy et al., 2018; Dobson, 2019). The theoretical foundation of distributional semantics is often traced back to J. R. Firth's celebrated dictum: "You shall know a word by the company it keeps" (Firth, 1957: 11). Deep learning models do not treat words as isolated symbols; rather, they model them within their distributional contexts, calculating the probability that particular terms occur in proximity to one another across vast textual corpora. Through this process, textual units such as words, sentences, or documents can be represented as dense vectors composed of numerical values (Mikolov et al., 2013; Pennington et al., 2014; Underwood, 2019). These vectors position texts within a multidimensional semantic space in which geometric proximity may be understood as reflecting relations of meaning (Turney & Pantel, 2010; Erk, 2012). Within this distributional framework, meaning itself is treated as a relative position in vector space, determined by the relationships that a term establishes with other elements of the corpus (Bengio, Courville, and Vincent, 2013).

Within this mathematical framework, meaning is conceived as a spatial position. For example, if two of Calvino's cities share the same melancholic atmosphere, similar architectural archetypes, or a comparable utopian tension, their corresponding vectors would tend to occupy nearby positions in semantic space. Proximity (typically measured through mathematical functions such as cosine similarity) may be understood as a computational approximation of semantic and stylistic affinity. Vector representations thus allow computational analysis to move beyond literal lexical correspondence. An algorithm can recognize that terms such as "mirror" and "crystal" are conceptually related even though their lexical forms are entirely

different. Such associations may make it possible to map deeper isotopies within the work with a degree of contextual sensitivity that was less accessible to earlier computational methods. The interpretive value of these mappings, however, depends on the critical framework in which they are situated.

The technological framework also requires clarification of the epistemological stance adopted in this study. The use of artificial intelligence in literary criticism should not be understood as a concession to scientistic positivism, nor as an attempt to delegate aesthetic judgment or the discovery of the text's "objective truth" to the machine. Indeed, drawing upon the criticisms advanced by Nan Z. Da (2019), Ciotti has highlighted the limitations of AI-driven approaches and cautioned against their uncritical application to literary analysis (Ciotti, 2021; 2023). Rather, AI is conceived here as akind of "interpretive collaborator," a dialogic partner to the human critic. The machine possesses a specifically non-human cognitive capacity: in our case, it can process, compare, and project all fifty-five cities simultaneously within vector space, allowing their relative positions to be considered across the work as a whole. Whereas conventional reading proceeds through a linear and diachronic sequence, AI-based analysis can examine the text's elements simultaneously within a shared vector space. It may detect patterns, structural anomalies, and emotional trajectories that human readers might intuitively perceive during close reading but find difficult to quantify or articulate on a macroscopic scale. Whether what the algorithm detects constitutes genuine pattern recognition or a sophisticated statistical artifact remains, however, an open methodological question. The task of interpreting the cultural and poetic significance of these patterns ultimately remains the responsibility of the human critic, Throughout the analysis, the interpretation of the data results from an ongoing methodological exchange between artificial intelligence and the human critic, one of prompting, verification, and refinement rather than delegation. While the final elaboration is entrusted to the author, this elaboration is never conducted in isolation from that exchange.

This convergence between computation and literature has a significant parallel in Italo Calvino's own intellectual trajectory. In the essay previously discussed, *Cybernetics and Ghosts*, Calvino prophetically envisioned the

emergence of a "literature machine": a device capable of combining predefined elements according to precise rules in order to generate narratives, challenging the Romantic myth of divine inspiration and emphasizing instead the artisanal and combinatorial dimension of literary production. As Calvino writes:

> In fact, given that developments in cybernetics lean toward machines capable of learning, of changing their own programs, of developing their own sensibilities and their own needs, nothing prevents us from foreseeing a literature-machine that at a certain point feels unsatisfied with its own traditionalism and starts to propose new ways of writing, turning its own codes completely upside down (Calvino, 1986: 11).

The passage defines literature as a combinatorial process and anticipates themes that have become central to contemporary Digital Humanities and generative artificial intelligence. This convergence between Calvino's combinatorial imagination and computational analysis provides a rationale for the AI-assisted approach adopted in this study. These theoretical premises are operationalized through a methodological pipeline comprising three stages, each corresponding to one of Franco Moretti's principal analytical tools: graphs, maps, and trees (Moretti, 2005).

2. Act I – Graphs: Quantitative Patterns and Emotional Space

The geometric structure of Invisible Cities appears to the reader as a crystalline mechanism; its emotional landscape, however, follows a wave-like dynamic that can escape the rigid sequence imposed by the table of contents. To explore these emotional fluctuations within Calvino's prose, this chapter adopts a quantitative sentiment analysis methodology supported by Large Language Models (LLMs). The data extraction process was carried out by submitting the entire corpus of the fifty-five cities (excluding the framing sections and the dialogues between Kublai Khan and Marco Polo) to the computational analysis of *Claude 3.5 Sonnet*, a model capable of working beyond lexical frequencies with the semantic density of literary prose. The model's outputs represent probabilistic interpretations rather than objective measurements, and their validity as literary evidence depends on the critical framework within which they are read. The analytical categories and scoring criteria were defined by the author in accordance

with the research questions and theoretical framework. The prompt used to operationalize these criteria was formulated with the assistance of *Claude 3.5* and subsequently reviewed and adopted by the author. The final prompt is presented below in its English translation from the original Italian version:

Step 1:

Act as an expert in Digital Humanities and Text Mining. I have uploaded the PDF version of Invisible Cities by Italo Calvino. Your task is to parse the text and extract all fifty-five cities. Create a structured table with the following columns:

1. *Sequential number (from 1 to 55)*
2. *Name of the city*
3. *Thematic category to which the city belongs (e.g., Cities and Memory, Cities and Desire, etc.)*
4. *Chapter in which the city appears (from 1 to 9)*
5. *Full text of the city's description*

For verification purposes, display only the first three lines of each description.

Step 2:

Now perform an in-depth analysis of the text of each of the fifty-five cities. For every city, assign three quantitative scores. Scores must be expressed with two decimal places and follow the scales specified below:

- *Emotional Polarity (Sentiment): from -1.00 (extreme negativity, anguish, death, or profound decay) to +1.00 (utopia, vitality, and lightness).*
- *Semantic Complexity: from 0.00 (purely material or objective description) to 1.00 (very high density of metaphors, paradoxes, and conceptual abstractions).*
- *Desire–Melancholy Tension: from -1.00 (oppressive nostalgia for the past) to +1.00 (forward-looking, future-oriented projection).*

Return the results exclusively in CSV table format, so that they can be easily copied into Excel (The reference structure of the table is provided in Appendix 1).

The emotional scale defined in the prompt, anchored at one end by "anguish, death, and profound decay" and at the other by "utopia, vitality, and lightness", already encodes a specific interpretive polarity. The scores subsequently generated by the model are therefore responses to this pre-structured framework rather than independent measurements of emotional

content. This chapter examines the work along its diachronic axis, following the linear sequence of reading from the first to the fifty-fifth city. Such a diachronic perspective makes it possible to explore whether the progression of the text conforms to a static architecture or is instead driven by an invisible thermodynamic current that gradually alters the emotional temperature of the work from one chapter to the next.

The diachronic visualization of emotional polarity across the fifty-five cities (Fig. 1) suggests that Calvino's text does not constitute a static or balanced atlas. The trajectory traced by the graph resembles a thermal curve, tracking a progressive shift in the work's emotional temperature.

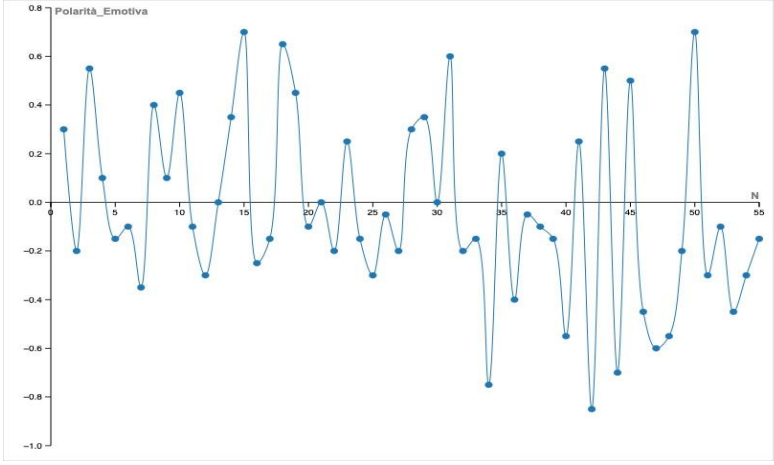


Fig. 1: Line Chart of Emotional Polarity

Observing the graph from left to right, a clear difference emerges at the macro-textual level between the first and second half of the journey. In the first twenty-five cities (roughly corresponding to chapters 1 to 4), the prose remains within a relatively controlled emotional oscillation. Positive peaks are frequent and reach significant heights (notably around cities 15 and 18), which can be associated with Marco Polo's sense of wonder in the presence of an urban repertoire still largely dominated by "lightness," utopian drive, and the plasticity of memory. At this stage, a form of equilibrium still

appears to hold; the empire, however vast, remains intelligible and geometrically manageable within the cognitive framework of the Great Khan.

A major macro-textual turning point occurs around city 30, coinciding with the theoretical midpoint of the book, where the graph undergoes a marked morphological shift. The oscillations of the line become more pronounced, producing abrupt jumps that resemble a seismic trace. It is at this stage that the graph registers the most striking collapses in the entire work: city 34 registers a sharp drop, followed shortly after by city 42, which reaches the lowest point in the entire series. From a hermeneutic perspective, these quantitative lows coincide with the emergence on the page of the "continuous cities" (Procopia, Cecilia, Pentesilea), metropolitan nightmares made of endless asphalt and chaotic accumulation, leading to the dissolution of identity.

The text's response to these collapses is of critical interest. Immediately after these emotional troughs, Calvino's prose reacts with sharp upward spikes (as can be seen between cities 42-43 and in the isolated peak of city 50). This pattern can be read as reflecting an effort to erect a barrier of rationality and geometric harmony (the airy or threadlike cities) against the advance of the formless. However, this resistance appears to be only temporary. After the final major peak at city 50, the line is no longer able to return to positive territory, remaining below zero through the final five urban units. The quantitative trajectory thus suggests a progressive collapse of the imperial system. The Khan's chessboard appears to be gradually exhausted; the computational analysis indicates that Invisible Cities may function not as a disordered collection of postcards, but as a controlled kinetic descent toward the realization of the urban inferno.

The diachronic analysis of the line has offered a provisional glimpse of the temporal rhythm of what may be an imperial collapse; the full heatmap (Fig. 2) now enables a synchronic exploration of Calvino's imaginative universe, isolating the apparent emotional coloration of the eleven thematic categories across all nine chapters of the work. The matrix organises the urban fragments by faithfully reproducing the combinatorial structure based on descending diagonals designed by the author, yet the chromatic spectrum produced by the algorithm suggests a potential point of rupture in

the textual system. The observation invites further critical scrutiny rather than constituting a definitive interpretive conclusion.

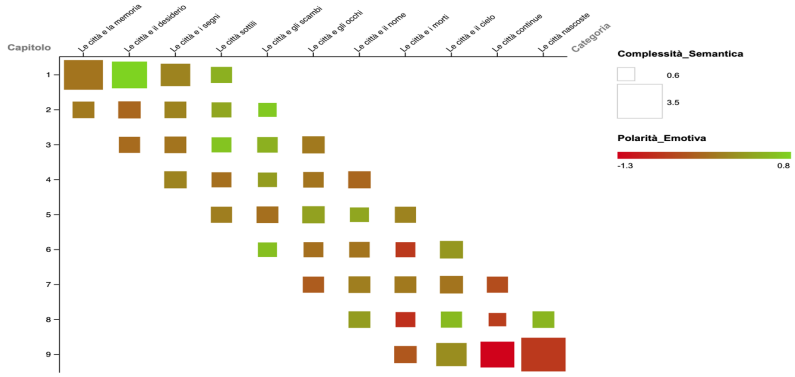


Fig. 2: Complete Category Heatmap

The category *The Cities and the Dead* might seem, at first glance, the emotional centre of Calvino's sense of unease, yet the quantitative data visualised in Fig. 2 challenge this impressionistic reading. The cells corresponding to the cities of the dead (distributed between chapters 5 and 8) do not in fact register the most acute negative polarity. Instead, they maintain a stable ochre-brown tone, warm and melancholic, but ultimately controlled. In Calvino's imagination, death, understood either as a mirror of memory or as a symmetrical and almost elegant dimension of the living (Eusapia, Laudomia), still retains a certain geometric dignity: a form of stasis that does not produce horror, but rather a detached philosophical contemplation. The epicentre of emotional collapse, the point at which the algorithm registers the coldest, most dystopian temperature across the entire work, concentrates with unexpected intensity in Chapter 9. It is in this final row of the matrix that the heatmap undergoes a dramatic chromatic shift, flaring into a dense, dark blood-red precisely in correspondence with the *Continuous Cities* and the *Hidden Cities*. Here, artificial intelligence isolates and geometrically quantifies the true traumatic repression at the core of Calvino's writing. What unsettles the author is not biological death itself (the cities of the dead), but rather the totalising alienation of modernity: the

formless expansion of the megacity that devours the planet (Cecilia, Pentesilea), and the suffocating condition of hidden cities in which unhappiness has become an invisible gas.

Alongside this terminal abyss, the heatmap also reveals, in contrast, the persistence of bright green oases confined almost exclusively to Chapter 1 (*Cities and Desire*), testifying to an initial utopian purity that gradually disintegrates across the book's progression. This precise chromatic transition from top to bottom demonstrates how AI can objectify the complex thermodynamics of the text. Calvino employs the geometric grid of the index to stage a gradual and relentless philological reckoning: beginning with the bright green of youthful desire, the prose descends through the shadowed terrains of memory and signification, before ultimately collapsing into the terminal red of the ninth chapter, the algorithmic marker of the contemporary urban inferno.

In conclusion, the patterns emerging from the analysis of the fifty-five cities suggest that the temporal structure of Calvino's narrative is anything but static. Its linearity appears to conceal a complex thermal fluctuation, oscillating between the green of utopia and the red of entropic alienation, though it should be acknowledged that these chromatic and emotional categories are themselves the product of the model's interpretive framework rather than intrinsic properties of the text. However, while the graphs appear to capture something of the temporal and categorical evolution of this polarity, they may lack the dimensional properties necessary to reveal the underlying conceptual proximities between individual cities beyond the constraints of chronology. Having explored the flow of emotional time through line-based visualisations, this chapter now turns to spatialising these conceptual distances: shifting from a chronological axis to a cartographic one, and moving toward the analytical framework of maps.

3. Act II – Maps: The Network of Latent Relationships

The index of Invisible Cities establishes a fixed order in which the cities are presented. The relationships between them, however, do not necessarily follow that sequence. The conversations between Marco Polo and Kublai Khan, together with recurring images and themes, invite connections across different groups of cities. This chapter explores these connections through

network analysis, representing the cities as nodes whose positions are determined by their semantic relationships rather than by their order in the book.

The resulting map is not an alternative geography of Calvino's world, but a representation of relationships produced by the analytical model, and therefore dependent on the way the texts are computationally represented and compared. This was the most complex stage of the analysis, since it required translating the interpretative framework into a procedure that could be applied consistently to each city. The analysis was therefore carried out in two stages. The analytical framework and the criteria for establishing relationships between the cities were defined by the author. *Claude 3.5 Sonnet* processed the individual city descriptions. The prompt itself was developed also with the assistance of *Claude 3.5*, from an Italian version prepared for the study and subsequently reviewed by the author. The two stages are described below. The prompt below follows the same Italian-to-English translation process described earlier.

Step 1:

Act as an expert in Digital Humanities and Network Analysis. Using the structured corpus of the fifty-five cities from Italo Calvino's Invisible Cities, your task is to identify latent relationships between the cities that are not made explicit by the book's index.

Analyse the descriptions and connect cities that share strong structural or symbolic affinities (for example: doubled or mirror cities, cities based on threads and networks, liquid cities, invisible or aerial cities, or cities associated with commerce or death).

Generate an edge list in CSV format with the following three columns:

- *Source: the name of City A*
- *Target: the name of City B*
- *Weight: the strength of the connection, expressed on a scale from 1 to 3*
    - *1 = weak or evocative affinity*
    - *2 = clear thematic affinity*
    - *3 = very strong structural affinity, approaching a form of twinship*

Select only the most significant relationships. Avoid connecting every city to every other city, and retain only links with weights of 2 or 3 in order to preserve the readability of the graph.

Return the results exclusively as a CSV code block.

Step 2:

Based on the edge list you have just generated, briefly describe the three or four main clusters (communities of cities) that emerge from the semantic network. Assign each cluster a critical and interpretative label (for example, Cities of Deceptive Doubling or Cities of Material Transience) and indicate which cities belong to each group (The reference structure of the table is provided in Appendix 2).

The categories suggested in the prompt, "doubled or mirror cities," "liquid cities," and "cities associated with commerce or death," are drawn from established critical approaches to Calvino's work. The clusters identified by the algorithm may therefore reflect, at least in part, the interpretive history of the text rather than structures that would emerge from a categorically neutral analysis.

The algorithm then analysed symbolic constants, philosophical orientations, and recurring metaphorical patterns, calculating the mathematical distance between the resulting vectors. The *weight* of the connections was derived from the resulting proximity matrix: the more closely two cities were related in terms of their speculative concerns, the stronger their semantic connection.

The resulting network was then projected and visualised in the graph-computing environment *Gephi*, with the fifty-five cities represented as nodes and their semantic relationships as edges. A force-directed layout algorithm was used to organise the graph. In this model, nodes repel one another in a manner comparable to charged particles, while edges function as springs that draw connected nodes closer together. The resulting graph translates the data's multidimensional structure into a two-dimensional representation that can be interpreted visually. Relationships that are not readily apparent in the linear sequence of the printed text thus become visible as spatial configurations, bringing out aspects of the underlying organisation of Calvino's imaginative universe.

The network produced by the algorithm (Fig. 3) provides a different visual configuration of Invisible Cities. It breaks with the continuity of the editorial sequence and redistributes the cities according to their semantic proximity. The resulting system is highly polarized, with a large central conceptual mainland and three isolated districts located at the periphery of the graph.

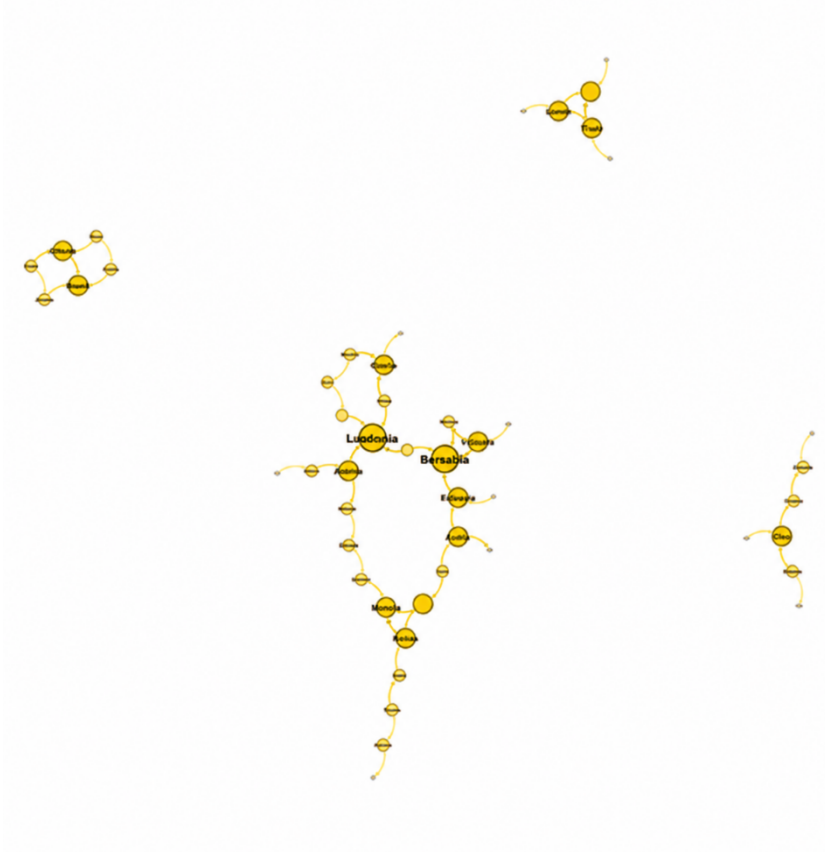


Fig. 3: Network graph of latent relationships

The core of the network coalesces around two major relational hubs, highlighted by the algorithm through larger and darker nodes: Laudomia and Bersabea. This central area of the network represents the speculative and metaphysical heart of the book. On the left, the connections extend from Clarice, Zora, and Maurilia and converge on a vertical axis centred on death and time. Laudomia, the city of the living, the dead, and the unborn, is closely connected to Adelma, the mirror-city of the deceased. From this point, the biological and social trajectory continues through Melania, Eutropia, and Sofronia, eventually reaching the group formed by Marozia and

Raissa, the cities of hidden unhappiness and human interconnectedness. Ipazia, Tamara, and Aglaura then form a further linear extension of this group.

On the right, the network is organised around Bersabea, the city that possesses both a celestial counterpart made of precious stones and an underground counterpart composed of refuse. This node brings together the book's cosmic and geometric imaginaries. Eudoxia, with its carpet-map, and Andria, the horoscope-city, are connected through Tecla directly to Marozia. Above Bersabea, forming a kind of crown around the central hub, are Valdrada and Moriana, both structured around reflection, mirroring, and visual inversion.

Taken together, this configuration suggests that artificial intelligence is grouping within a single macro-cluster several of Calvino's recurring obsessions: geometric symmetry, the relationship between life and the afterlife, the deceptive power of signs, and the search for cosmic order in the face of chaos. Whether these groupings reflect genuine thematic affinities or instead derive partly from the model's own representational biases remains an open question. One of the most striking features of the network shown in Fig. 3, however, is the position of the cities that the algorithm appears to have pushed towards the margins, reducing their connections with the central cluster. Calvino's imaginative universe may thus appear discontinuous through this computational process, structured around what the model interprets as distinct stylistic and conceptual departures. It remains unclear whether these discontinuities belong to the work itself or emerge instead from the embedding and projection process. The upper-left portion of the network contains a self-contained and highly cohesive group formed by Ottavia, the spiderweb city suspended above the abyss, Isaura, the city of a thousand wells, Baucis, Armilla, Ersilia, and Zenobia. These are the "thin cities." Their shared aerial, liquid, and filament-like qualities distinguish them from the material logic governing much of the rest of the urban universe and isolate them within an enclave of pure lyrical lightness.

A small and relatively isolated group appears in the upper right, where Leonia, the city of waste, is associated with Trude, the endlessly replicated city-copy. Their marginal position suggests that the nightmare of industrial modernity and consumer accumulation constitutes an anomalous body

within the broader symbolic economy of the work. It is the trauma of the formless, unable to reconcile itself with the crystalline order that characterises the rest of Calvino's cities. On the opposite edge of the network, the node centred on Cloe, the city of encounters and silent interactions, forms a separate group together with Zemrude, Despina, and Euphemia. Here the focus falls on merchants, travellers, and stories exchanged around the fire. The algorithm highlights the distinctive nature of these cities: they are not organised around architecture or geometry, but around circulation, communication, and the movement of words.

The map presented in Fig. 3 therefore offers a computational view of a Calvino who appears less as the weaver of a single homogeneous tapestry than as an architect of differentiated conceptual spaces, at least as shaped by the model's representational choices. At the centre of the network lies what may be a vast area devoted to philosophical and existential reflection, while the more experimental groups around it explore the limits of matter, seriality, and language. In this sense, the AI-based analysis offers one way of visualising how the invisible threads of memory, utopia, and alienation might connect chapters separated by dozens of pages in the printed text.

The network map nevertheless remains fundamentally synchronic. It indicates where concepts may be located in relation to one another, but it does not explain how these configurations might emerge from a broader morphological or genealogical perspective. The cartography of the network thus provides a provisional view of the spatial proximity of the cities within the semantic landscape, while leaving the formal and conceptual genealogy underlying these apparent affinities to further critical analysis. The next step, therefore, moves from the synchronic logic of the map to the evolutionary diachrony of the tree, exploring what might constitute the morphological phylogeny of Calvino's urban archetypes.

4. Act III – The Trees: Phylogeny and Mutation of the Urban Archetype

The network map has offered a provisional view of the synchronicity of semantic attractions within latent space; the Tree now explores the possible formal structure and morphological affinities of Calvino's imaginative

universe. Drawing on Franco Moretti's morpho-evolutionary approach, this chapter treats the fifty-five cities as formal configurations whose affinities can be explored through a provisional phylogenetic model. Moretti's approach, directly inspired by Charles Darwin's biological theories, applies the concepts of diversification and selection to the anatomy of texts: literary genres and styles may evolve by branching from common ancestors through minimal formal variations and expanding into specific morphological niches.

This model offers a way of approaching Calvino's literary laboratory. Rather than constructing fifty-five entirely self-contained entities, Calvino might be understood as introducing morphological and combinatorial variations on a single archetypal macro-theme. *Invisible Cities* may thus share a common underlying "biological" core while diversifying through trajectories of stylistic branching. The Tree makes these recurring patterns of formal variation visible, turning the book's combinatorial index into a provisional morphological map in which vertical and Euclidean distance may suggest degrees of formal and structural affinity between the individual chapters.

To map this genealogical branching mathematically, the "textual genes" were encoded with the assistance of *Claude 3.5*. The same model was used to generate the prompt for the present analysis, which was subsequently reviewed and adopted by the author. The prompt itself is given in English below, translated from the original Italian. The analysis proceeded in two distinct stages:

Step 1:

Act as an evolutionary biologist applied to literary criticism (Digital Humanities). Your task is to map the stylistic "genetic profile" of the 55 cities and the 9 dialogic frames in Italo Calvino's Invisible Cities for a Hierarchical Clustering analysis.

For each of the 55 city-descriptions and the 9 dialogic frames, analyse the text and assign a quantitative value (on a scale from 0 to 5, indicating intensity) to the following six stylistic "genes":

- *Sintassi_Cristallina (Crystalline Syntax): the degree to which the prose is organised through clear, geometric, visually precise sentence structures. (0–5)*

- *Sintassi_Accumulo (Accumulative Syntax): the degree to which the prose proceeds by accumulation, enumeration, or cumulative listing of objects, actions, or images. (0–5)*
- *Presenza_Io_Narrante (Narrating "I" Presence): whether an explicit first-person narrating voice addresses the Great Khan or intervenes directly in the text. (0 = absent, 1 = present)*
- *Dimensione_Metanarrativa (Metanarrative Dimension): whether the text reflects on its own status as narration, description, or act of storytelling. (0 = absent, 1 = present)*
- *Frizione_Descrittiva (Descriptive Friction): the degree of internal paradox, resistance, or estrangement between what the city appears to be and what it is revealed to be. (0–5)*
- *Tensione_Aforismatica (Aphoristic Tension): the degree to which the passage closes on, or is structured around, an aphoristic, sententious, or maxim-like statement. (0–5)*

Return the data strictly in CSV format or as a Markdown table, one row per unit, ordered as in the source text (M-1 to M-5, De-1 to De-5, Se-1 to Se-5, So-1 to So-5, Sc-1 to Sc-5, Oc-1 to Oc-5, N-1 to N-5, Mo-1 to Mo-5, Ci-1 to Ci-5, Co-1 to Co-5, Na-1 to Na-5, and D-I to D-IX for the dialogic frames).

Step 2:

Based on the resulting data matrix, write Python code to be executed in Google Colab using the scipy.cluster.hierarchy library. The code should compute distances between cities and generate a dendrogram (hierarchical tree).

Display city names as leaf labels on the tree and use either the "ward" or "complete" linkage method. Color the branches according to the identified macro-clusters (The reference structure of the table is provided in Appendix 3).

The six stylistic parameters used as "genetic markers" constitute an a priori critical taxonomy defined by the researcher. The clusters that emerge from the analysis therefore reflect this particular coding system rather than fixed or objectively given literary categories. A different selection of parameters would produce a different dendrogram: the phylogenetic tree presented here depends on the analytical categories established at the outset.

The fifty-five city-descriptions and nine dialogic frames were examined for six recurring formal and stylistic parameters: the degree of crystalline, geometrically ordered syntax; the degree of accumulative, enumerative syntax; the presence of an explicit narrating "I" addressing the Great Khan; the presence of a metanarrative dimension reflecting on the act of storytelling;

the degree of descriptive friction, understood as the internal paradox or estrangement between a city's appearance and its revealed nature; and the degree of aphoristic tension in the passage's closure. Each unit was consequently coded as quantitative data.

These features were then used for hierarchical clustering according to Ward's method and Euclidean distance. The procedure is agglomerative: each city initially constitutes an independent unit, and cities or groups of texts are progressively merged according to the variance in their genetic profiles. The resulting dendrogram (Fig. 4) arranges the texts according to their formal affinities. Its branching structure records the successive stages of statistical aggregation, while the terminal branches retain the individual stylistic profiles of the texts.

At the distance threshold indicated by the red dashed line in Fig. 4, the dendrogram yields five major morphological clusters, which organise the formal variation found across the text.

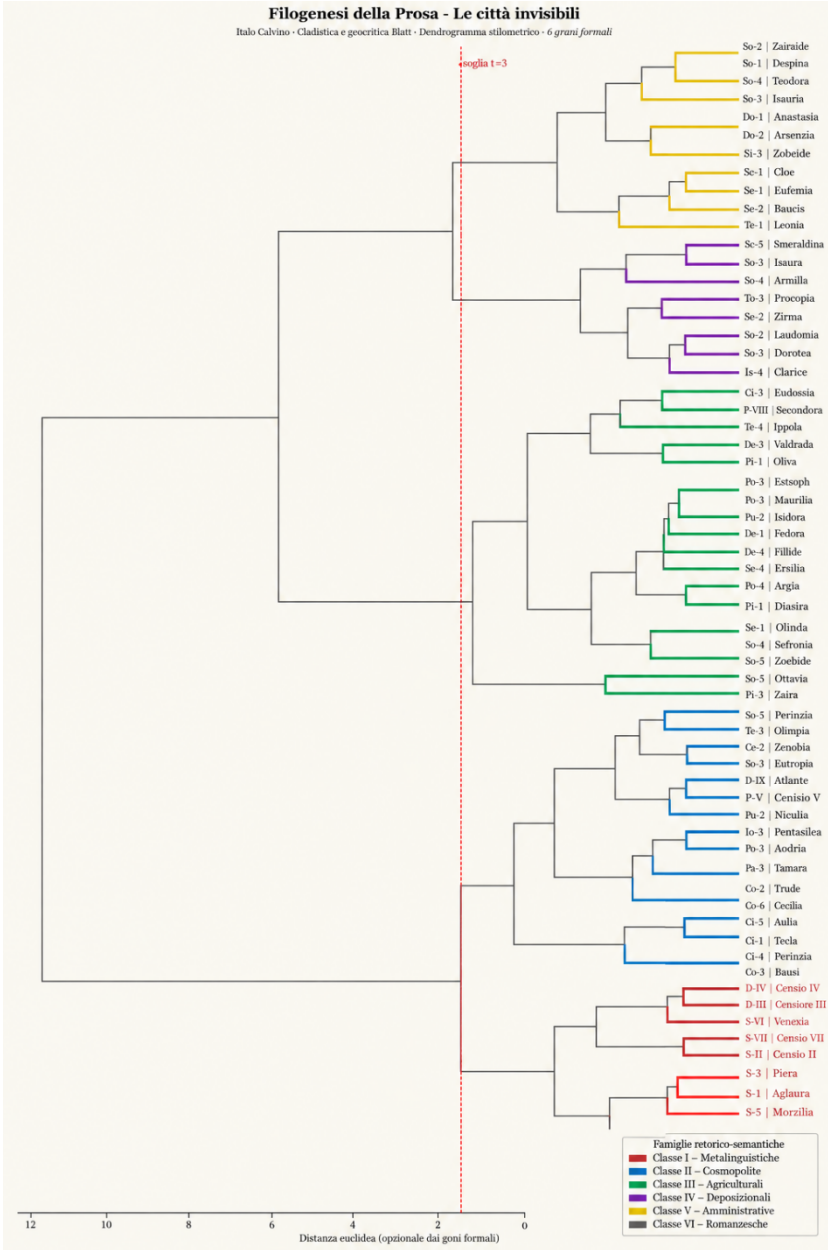


Fig. 4: Dendrogram of the Phylogeny of Prose

The root structure shows a first major bifurcation of the common trunk. One branch brings together Clades I and IV, corresponding respectively to metalinguistic and aphoristic tendencies; the other comprises Clades II, III, and V, where accumulative, objectual, and nominalistic tendencies converge. This division need not be understood as a fixed opposition. It instead points to two broad configurations of formal organisation within Calvino's prose.

At the level of the individual leaves, that is, the urban texts, the algorithm identifies possible genetic affinities among the smallest branches. Diomira (M-1) and Isidora (M-2), for instance, are closely related within Clade III. The case of Zenobia and Despina is more striking: the two cities receive identical scores across all six dimensions and therefore occupy the same position in the quantitative representation. Their numerical profiles coincide in several respects. Both score highly for crystalline syntax and descriptive friction, receive no score for narrating-I presence or metanarrative dimension, and show similarly high levels of aphoristic tension across all six dimensions. The clustering may thus be identifying a recurring combination of formal precision, descriptive friction, and aphoristic closure.

The Venice hypothesis, already noted by Ossola (1987), Barenghi (1992), and Segre (1994), is particularly significant in this context. Within Clade I (Red), D-VI Venice is grouped with several dialogic and metanarrative frames, including Frames II, III, IV, and VII, and belongs to the same broader cluster as Pirra, Aglaura, and Maurilia. The configuration recalls Marco Polo's well-known confession to the Great Khan that every time he describes a city, he is saying something about Venice (see "Every time I describe a city I am saying something about Venice", Calvino, 2023: 78). The correspondence remains an interpretive resonance rather than a proof.

Within the coding scheme adopted here, Venice is therefore associated primarily with the metalinguistic configuration of Clade I, rather than with the accumulative configuration of Clade II or the strongly descriptive configuration of Clade III. The "gene" of Venice, understood as a possible reflection on language, absence, and memory, resonates with the dialogic dimension of the work. Venice is not a literal genealogical origin of the other cities; the clustering results instead offer an intriguing formal parallel to its symbolic function as a possible metalinguistic matrix.

The bio-stylistic analysis represented by the dendrogram in Fig. 4 offers experimental indications that Invisible Cities constitutes a coherent and structurally organised phylogenetic ecosystem. The hierarchical clustering suggests that the work's apparent fragmentary nature corresponds to the cyclical evolution of five formal families of prose, with each text functioning as a regulated variation within a single overarching literary framework. These results remain hypotheses rather than empirical demonstrations, given the limitations of the model and the interpretive steps involved in moving from mathematical clustering to literary conclusions.

The phylogenetic tree maps the formal affinities identified across the six selected stylistic dimensions. Within the coding system used here, the recurrent configurations and morphological affinities can be described through the metaphor of phylogenetic branching. The multiplicity of Calvino's cities can thus be approached as a system of recurring and differentiating formal patterns. This third analytical instrument brings the experimental investigation of Calvino's geometry to a provisional conclusion. It leads, finally, to a consideration of the relationship between artificial computation and literary humanism, and of the methodological lessons this exploratory exercise may offer for the future of AI-assisted literary criticism.

5. Conclusions: An Artificial Humanism?

At the end of this computational exploration of Italo Calvino's work, the methodological triad of Distant Reading, graphs, maps, and trees, points toward what this study calls Artificial Humanism. The aim here is to use AI-assisted methods to investigate possible structures within the text, without reducing the lyrical density of *Invisible Cities* to statistical data. Working in a vector space of meanings, AI has identified configurations of ideas, structures, and anxieties that may not be immediately apparent in linear reading. Whether such configurations amount to interpretive insight or remain instances of statistical patterning cannot, however, be determined by the computational analysis alone.

The three procedures provide a provisional picture of the novel's semantic and formal organization. The line chart and heatmap show a marked intensification of semantic fluctuation in the second half of the book. The strongest tension emerges in metropolitan and industrial alienation, well beyond

the relatively neutral register of biological finality found in Cities and the Dead. The network graph places Laudomia and Bersabea at the centre of the relational structure, with satellite groupings such as Trude and Leonia, while the dendrogram divides the fifty-five cities into five stylistic families. The proximity between Venice and the dialogic frames is particularly suggestive of a possible generative logic. These mathematical relations remain projections within a reduced dimensional space rather than direct properties of the source text.

Epistemological caution is therefore part of the inquiry itself. As Fabio Ciotti (2021) argues, the application of computational language models to literature should first be understood as an empirical test of their actual critical and interpretive capacities. The analysis encounters here the structural limitation of the "black box": the vectors and probabilistic operations through which the system identifies apparent geometries in Calvino's text remain largely opaque to the humanist scholar. The AI's retroactive explanations cannot consequently serve as independent theoretical justification for the patterns it produces. The scholarly value of the experiment lies instead in the critic's ability to interrogate these outputs while remaining aware of the computational processes that produced them.

The relationship between computational and humanistic culture is also at stake. In *Cybernetics and Ghosts*, Calvino imagined a "literary machine" capable of revealing the rules underlying literary creation. AI-driven Distant Reading can be understood as a contemporary extension of that possibility, providing new ways of testing the formal and mathematical dimensions of literary texts. It gives the critic another vantage point from which the text can be observed and questioned, another seat, perhaps, at the game Marco Polo and Kublai Khan play across the book's dialogic frames, without replacing the critic's interpretive work. The experiment thus contributes, however modestly, to the possibility of treating scientific-mathematical and humanistic approaches as complementary modes of inquiry rather than opposing domains.

The experiment presented here is a case study in the possibilities and tensions that arise when computational systems enter the hermeneutic space traditionally occupied by human readers, rather than a search for definitive truths about *Invisible Cities*. Its object of investigation is therefore not only

Calvino's text, but also the forms of interpretation that emerge from the encounter between literary criticism and generative AI. Artificial Humanism, in this context, does not designate the replacement of human interpretation by computational analysis. It names a methodological space in which computational patterns become material for renewed critical interrogation.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

During the preparation of this article, the author used generative AI-assisted tools at different stages of the research process. These tools were used to assist with literature searches, the verification of bibliographic records against publisher and citation databases, the extraction and structuring of textual data, and the linguistic and stylistic revision of author-generated text. Generative AI was also used as an analytical instrument to process the corpus, assign quantitative scores according to categories established by the author, identify semantic relationships, and generate data for network analysis and hierarchical clustering. The categories, coding criteria, research questions, and interpretive framework were defined by the author, who critically verified and interpreted the model-generated outputs. AI tools did not replace scholarly judgment or provide autonomous literary interpretation. The author remains responsible for the evidence, analysis, theoretical argument, and conclusions, and takes full responsibility for the content of the publication.

Bibliography

Aragona, R. (2008). *Italo Calvino. Percorsi potenziali*. Manni.

Balice, M. (1986). Le città di Calvino. *Paragone Letteratura*.

Barenghi, M. (1992). Note e notizie sui testi. *Le città invisibili*. In I. Calvino, *Romanzi e racconti* (Vol. 2, pp. 1359-1365). Mondadori.

Barenghi, M., Canova, G., & Falcetto, B. (Eds.). (2002). *La visione dell'invisibile. Saggi e materiali su «Le città invisibili» di Italo Calvino*. Mondadori Electa.

Battistini, A. (2001). Le città visibili e invisibili di Italo Calvino. *Esperienze letterarie, 26*(2), 21-38.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(8), 1798-1828.

Bloom, H. (2000). Italo Calvino [*Le città invisibili*]. In *Come si legge un libro (e perché)* (pp. 71-75). Rizzoli.

Brera, M. (2011). At the court of Kublai Kan: Storytelling as semiotic art in *Le città invisibili* by Italo Calvino. *Symposium: A Quarterly Journal in Modern Literatures, 65*(4), 271-289.

Calvino, I. (1972). *Le città invisibili*. Einaudi.

Calvino, I. (1986). *The uses of literature*. Harcourt Brace & Company.

Calvino, I. (2023). *Invisible cities* (W. Weaver, Trans.). Vintage Classics. (Opera originale pubblicata nel 1972)

Ciotti, F. (2021). Distant reading in literary studies: A methodology in quest of theory. *Testo e Senso, 23*, 195-213.

Ciotti, F. (2023). Minerva e il pappagallo. IA generativa e modelli linguistici nel laboratorio dell'umanista digitale. *Testo e Senso, 26*, 289-315.

Da, N. Z. (2019). The computational case against computational literary studies. *Critical Inquiry, 45*(3), 601-639.

Dobson, J. E. (2019). *Critical digital humanities: The search for a methodology*. University of Illinois Press.

Erk, K. (2012). Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass, 6*(10), 635-653.

Firth, J. R. (1957). A synopsis of linguistic theory 1930–1955. In *Studies in Linguistic Analysis* (pp. 1-32). Blackwell.

Hu, H., & Amaral, P. (2022). Word embeddings and semantic shifts in historical Spanish: Methodological considerations. *Digital Scholarship in the Humanities, 37*(2), 441–461.

Jänicke, S., Franzini, G., Cheema, M. F., & Scheuermann, G. (2017). Visual text analysis in digital humanities. *Computer Graphics Forum, 36*(6), 226-250.

Jockers, M. L. (2013). *Macroanalysis: Digital methods and literary history*. University of Illinois Press.

Kuon, P. (1987). Utopie-Kritik und Utopie-Entwurf in «Le città invisibili» von Italo Calvino. *Italienische Studien, 10*.

Leavy, S., Gleeson, J., & Greene, D. (2018). Navigating literary text with word embeddings and semantic lexicons. In *Proceedings of COMHUM 2018*.

Lima, E. (2023). *Le città invisibili* as cybertext and cyberspace: Italo Calvino, Ted Nelson, and Arata Isozaki. *Italian Studies, 78*(2), 192-209. https://doi.org/10.1080/00751634.2023.2220538

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space*. arXiv.

Moretti, F. (2000). Conjectures on world literature. *New Left Review, 1*, 54-68.

Moretti, F. (2005). *Graphs, maps, trees: Abstract models for a literary history*. Verso.

Moretti, F. (2013). *Distant reading*. Verso.

Ossola, C. (1987). L'invisibile e il suo "dove": «Geografia interiore» di Italo Calvino. *Lettere italiane, 39*(2), 220-251.

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532-1543).

Ravazzoli, F. (1987). «Le città invisibili» di Calvino: Utopia linguistica e letteratura. *Strumenti critici, 2*(2), 193-201.

Romele, A., Severo, M., & Furia, P. (2020). Digital hermeneutics: From interpreting with machines to interpretational machines. *AI & Society, 35*(1), 73-86.

Segre, C. (1994). Italo Calvino. «Le città invisibili» e la vertigine epistemica. *Strumenti critici, 19*(104), 43-53.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research, 37*, 141-188.

Underwood, T. (2017). A genealogy of distant reading. *Digital Humanities Quarterly, 11*(2).

Underwood, T. (2019). *Distant horizons: Digital evidence and literary change*. University of Chicago Press.

Wevers, M., & Koolen, M. (2020). Digital Begriffsgeschichte: Tracing semantic change using word embeddings. *Historical Methods, 53*(4), 226-243.

Wellner, G. (2020). Material hermeneutic of digital technologies in the age of AI. *AI & Society*.

Zancan, M. (1996). *Le città invisibili* di Italo Calvino. In *Letteratura italiana. Le opere* (Vol. 4, Parte 2, *La ricerca letteraria*). Einaudi.

Generative AI-Mediated Reception of Literary Texts in French as a Foreign Language: A Six-Stage Pedagogical Framework

Şule Çinar Yağci[1]

1. Introduction

Literary texts in foreign language education are not merely materials for teaching vocabulary and grammar. They enable learners to encounter language in context, construct meaning, interpret texts, and develop different perspectives. With the spread of the communicative approach, literary texts have therefore regained importance in foreign language classrooms, since genuine communication takes place through coherent texts whose meaning emerges within context rather than through isolated sentences (Ünal, 2005, p. 204–205). They also bring authentic language use into the classroom, an element often only partially represented in coursebooks (Bulut, 2017). In this sense, literature classes support not only language skills, but also cultural awareness, interpretation, and critical thinking (Bloemert et al., 2019, p. 377–380). Thus, the literary text should be approached as a multidimensional learning space that connects language, culture, and meaning.

This issue is especially significant in the teaching and learning of French as a foreign language, since French language teaching involves not only grammar, vocabulary, and communicative patterns, but also the cultural, historical, and aesthetic heritage of the French and Francophone worlds. Studies on French coursebooks used in Türkiye show that cultural elements and artistic signs hold an important place in these materials (Demirkan, 2020; Kuşçu, 2020). Demirkan's (2020) document analysis of *Le Nouveau Taxi 1 Méthode de Français* found that the sociocultural elements embedded in the instructional

[1] Lecturer Dr., Zonguldak Bülent Ecevit University, School of Foreign Languages, ORCID: 0000-0003-4456-3174, *sule.yagci@beun.edu.tr*, Türkiye

content and supported through dialogues were appropriate for A1-level objectives and sufficient for transmitting the target culture. Similarly, Kuşçu (2020, p. 647) emphasizes that knowing a foreign language requires not only mastery of grammar and vocabulary, but also familiarity with the culture conveyed through that language. However, the mere presence of cultural elements in coursebooks does not ensure their meaningful treatment. Therefore, French classes should focus not only on including literary and cultural texts, but also on how students understand, interpret, and integrate them into their learning process.

What makes literary texts pedagogically valuable is that they cannot be reduced to a single correct meaning. According to Iser's (1972, p. 281–285) classical approach, the text is not a complete and closed structure independent of the reader; rather, it gains meaning through the reader's expectations, interpretations, responses, and ways of filling textual gaps. This view provides an important theoretical basis for foreign language teaching. Ünal (2005, p. 206–209) similarly argues that directing students toward a single interpretation predetermined by the teacher weakens the communicative nature of the lesson. Literary reading in the foreign language classroom should therefore not be treated as the search for a fixed meaning. Instead, it should be seen as a dynamic learning experience that reveals the student's relationship with the text, interpretive choices, and meaning-making process. Through this process, students make inferences, complete textual indeterminacies, compare interpretations, and question their own reading. Literary reading thus becomes a multidimensional practice that supports linguistic, intellectual, cultural, and critical development.

Generative artificial intelligence (AI) tools offer new possibilities for foreign language teaching. Recent studies show that these tools can support personalized interaction, immediate feedback, explanation, reformulation, and example generation. In Slamet's (2024) study, teachers and students stated that ChatGPT could contribute to language acquisition, motivation, and individualized learning. Mahapatra's (2024) mixed-method intervention study also showed that ChatGPT can support students' writing development when used for instructional feedback. However, generative AI also entails risks, particularly regarding output accuracy, insufficient contextual understanding, source reliability, and overdependence on the tool. Zhai and colleagues (2024) warn

that unverified AI outputs, when used without critical evaluation, may weaken decision-making, critical thinking, and analytical reasoning. Therefore, the role of generative AI in foreign language teaching should be evaluated not only through its technical benefits, but also through its effects on students' thinking, questioning, interpretation, and learning responsibility.

The central question is whether generative AI will deepen literary reading as a support tool or reduce students' reading and thinking effort by directing them toward ready-made interpretations. The answer depends less on the presence of the tool than on its pedagogical purpose and mode of use. As Hirvela (1996, p. 127–128) notes, allowing students' personal responses to remain entirely unrestricted does not necessarily lead to deep reading; it may instead produce superficial interpretations detached from the text. Reader-response pedagogy becomes meaningful only when the student's initial response is transformed into a more text-based, questioning, and justified interpretation. The same principle applies to generative AI. If the tool is positioned as an authority that thinks and interprets on behalf of the student, the student's interpretive responsibility weakens and literary reading may become passive. Yet when used within a gradual, supervised, and purposeful reading process planned by the teacher, generative AI can help students analyze the text more carefully. In this way, it may serve as a pedagogical support tool for identifying layers of meaning, comparing interpretive possibilities, discussing ambiguities, and reconsidering incomplete or inaccurate interpretations.

This study aims to establish a theoretical and instructional basis for the reception of literary texts in teaching French as a foreign language, and to discuss the conditions under which generative AI may contribute to this process. It focuses on the function of literary texts in foreign language teaching, the instructional significance of reception, the difficulties encountered in French language education, and the opportunities and limitations of generative AI. It also explains the need for a new, balanced, and pedagogically controllable framework for working with literary texts. Accordingly, the following sections first examine the place of literary texts in foreign language teaching, then discuss the reception processes of learners of French. The study then evaluates the contributions and risks of generative AI in literary reading, and finally presents the rationale for a framework that brings together teacher guidance, student agency, critical reading, and AI use in a balanced way.

2. The Place of Literary Texts in Foreign Language Teaching

Literary texts are important instructional resources in foreign language teaching because they present authentic language use within coherent contexts. Genres such as poetry, short stories, novels, and drama enrich the classroom not only with vocabulary, but also with different modes of expression, stylistic layers, and context-dependent uses of language (Bulut, 2017, p. 4–5). They therefore help students encounter language as a system shaped by meaning, discourse, and culture. According to Ersan (2019, p. 289–290), learners who are geographically and socially distant from the target language often experience its natural use through literary and short texts. This shows the value of literary texts for cultural encounter and linguistic awareness. Similarly, Bloemert and colleagues (2019, p. 378–380) found that students largely associate literature lessons with vocabulary, idioms, and language skills. For this reason, literary texts should not be treated as merely supplementary materials. They are central instructional resources that reveal the workings of language in an aesthetic, condensed, and culturally rich form. Hall (2015) likewise argues that literature offers discursive, interpretive, and cultural possibilities that ordinary coursebook texts often cannot provide. Literary texts thus help students understand how language constructs narrative, perspective, evaluation, temporality, and social meaning.

Reducing literary texts to linguistic input narrows their pedagogical function. As Tilbe states, "each literary work acquires meaning through a set of distinctive linguistic, semantic, formal, and structural qualities" (Tilbe, 2019, p. 151). Literary texts support not only vocabulary, structure, and grammar learning, but also interpretation, meaning-making, self-development, and cultural awareness. Tsang, Paran, and Lau (2023, p. 1120–1121) show that students relate short stories, poems, and songs to language development as well as general skills, professional preparation, and personal growth. This confirms that literary texts are multidimensional learning resources in the foreign language classroom. Demirkan's (2020, p. 549–553) analysis of French coursebooks similarly shows that poems and texts about artists do not present culture as isolated information. Rather, they convey cultural meaning through linguistic, aesthetic, and symbolic elements. In this way, language, culture, and aesthetic form appear as interconnected

dimensions. The value of the literary text lies in this unity: it enables students to notice not only what language says, but also how it says it.

The literary text creates a multilayered field of meaning that requires the reader's active participation. According to Iser (1972, p. 281–287), textual gaps, silences, and implied meanings are completed through the connections established by the reader. Literary reading is therefore not simply a process of obtaining information from the text. Readers develop expectations, return to earlier parts, test possible interpretations, and revise them when necessary. This process helps students build a more conscious, questioning, and active relationship with the text. Yet, as Hirvela notes (1996, p. 127–128), personal response alone is not enough; it must be justified through textual evidence and transformed into critical reflection. Reader-response-based tasks can support students' search for meaning and foster higher-order thinking skills (Garzón & Castañeda, 2015, p. 187–188). From this perspective, the literary text should be seen as a learning space that develops meaning-making, inference, interpretation, and awareness of one's own thinking process.

The meaning-making function of literary texts is also closely connected to students' affective and intellectual development. Bloemert and colleagues (2019, p. 378–381) show that students do not expect literature classes only to provide historical or social context. They also value them for critical thinking, personal development, and self-expression. This confirms that literary texts broaden the learning experience in foreign language classrooms. Ünal (2005, p. 205–209) similarly emphasizes the importance of creative and subjective work in foreign language teaching. For Ünal, learning becomes more effective when students' personalities, interests, and experiences are included in the process. Therefore, the literary text should not be seen merely as a difficult text to be understood. It should be approached as a space in which students interpret, question, and reconstruct the relationship between themselves and the outside world. In this sense, literary texts help students become not only more competent language users, but also more attentive, questioning, and sensitive readers.

However, working with literary texts does not automatically produce effective learning. The outcome depends largely on how the text is handled in the classroom. As Ünal (2005, p. 206–207) notes, one common problem is

that teachers may guide students toward a single predetermined "correct interpretation." This limits students' ability to develop their own ideas and weakens creative participation. The lesson then becomes a restricted question-and-answer activity rather than an open process of meaning-making. Another problem is the selection of texts that do not match students' age, interests, or language level. In such cases, the literary text may become difficult, tiring, and discouraging rather than supportive of learning (Bulut, 2017, p. 8–9; Ersan, 2019, p. 289–290).

The presence of cultural elements in literary texts is likewise insufficient on its own. The way these elements are presented, supported through tasks, and discussed in class directly shapes the quality of learning (Demirkan, 2020; Kuşçu, 2020). Thus, the main issue often lies not in the literary text itself, but in why it is selected, how it is questioned, and within which instructional framework it is used. Literary texts bring together linguistic, cultural, intellectual, and affective dimensions; yet this richness requires careful planning. If text selection, task design, question types, and students' interpretive space are not properly organized, students may find the text too difficult or treat it as closed content with a single correct answer. Therefore, the value of literary texts in foreign language teaching can be realized only through a clear, guiding, and student-centered reading framework.

3. The Reception of Literary Texts in Teaching French as a Foreign Language

Reception rejects the view that textual meaning is fixed and already contained within the text. A literary text is an open structure that gains meaning through the reader's participation. According to Iser (1972, p. 281–285), readers complete what is not directly stated or is deliberately left open through their expectations, associations, and previous reading experiences. Reading is therefore not a passive act of receiving information, but an active and creative process of meaning-making. In foreign language teaching, Ünal (2005, p. 207–208) similarly shifts the focus of reception aesthetics from "What does the text mean?" to "What does the reader understand from the text?" This approach prevents the literary text from being treated as closed content explained by the teacher and merely confirmed by the student. Instead, the learner becomes an active reader who engages with the

text, makes inferences, and develops an interpretation. Interpretation is thus not a deviation from meaning, but a central element that deepens learning.

A student reading in a foreign language does not merely decode vocabulary and grammar. The student approaches the text through personal experience, the mother tongue, cultural memory, and previous reading habits. For this reason, literary reading becomes not only a linguistic process, but also a personal, cultural, and intellectual act of meaning-making. Garzón and Castañeda (2015, p. 187–188) state that reader-response-based practices position students as active subjects in search of meaning and develop higher-order thinking skills. Carlisle's (2000, p. 12–13) reading journal practice similarly reveals what students understand, what questions they ask, where they hesitate, and how they respond emotionally to the text. This is particularly important in teaching French, where the student engages with the literary text both as a foreign language user and as an interpreting reader. Therefore, the student's interpretation should be evaluated not by linguistic or philological perfection, but by the extent to which it is text-based, coherent, and justified.

When learners of French encounter a literary text, they interpret it on several levels at once. The first level is linguistic: students work through vocabulary, tenses and moods, forms of expression, associations, and syntax. According to Topaloğlu (2023), syntax is a system of rules that determines the correct and appropriate ordering of words in a sentence and explains how different sentence structures are formed. Yet literary meaning is not limited to linguistic analysis. The second level is cultural, involving proper names, places, social relations, ways of life, and historical references. The third level is aesthetic and interpretive, where students begin to notice point of view, silences, irony, rhythm, and plurality of meaning. As Demirkan (2020, p. 549–553) shows, when poetry, the figure of the artist, urban life, and everyday culture are presented together in French coursebooks, students do not simply acquire information; they also enter a field of cultural discourse. Reception in French language teaching therefore involves linguistic decoding, interpretive engagement, and the production of new discourse. As stated in the Common European Framework of Reference, reception is the process of receiving and processing input; in this process, learners activate appropriate schemata to construct a mental representation

of meaning and to form a hypothesis about the communicative intention behind it (Avrupa Konseyi, 2021, p. 51).

For foreign language learners, reception also entails certain difficulties. Limited language proficiency and distance from the target culture may cause students to experience a kind of "shock effect" when faced with a literary text (Ünal, 2005, p. 206–207). In contexts where daily contact with the target-language society is limited, the literary text may be both engaging and challenging (Bulut, 2017, p. 8–9; Ersan, 2019, p. 289–290). This difficulty is especially visible in French language teaching, since French literary texts may include dense cultural references, implicit allusions to historical events or places, and varied forms of expression. Such elements can slow down students' reception. Visual and sociocultural supports may make the text more accessible (Demirkan, 2020; Kuşçu, 2020), but their effectiveness depends on how the text is handled in class. Thus, reception difficulties cannot be explained only by students' linguistic limitations. Text selection, activity design, question quality, and teacher guidance are equally decisive.

In a French class, the teacher who supports reception does not lead students toward predetermined conclusions. Rather, the teacher organizes and deepens the interpretive relationship students establish with the text. Hirvela's distinction (1996, p. 127–128) is important here: a personal response is valuable, but it must be connected to textual clues. Carlisle's (2000, p. 12–13) reading journal practice and the reader-response-based tasks proposed by Garzón and Castañeda (2015, p. 187–188) likewise show that teachers can design activities that make students' interpretations visible. In this approach, questioning is not limited to "What did the author mean here?" Students are also asked, "What do you see in this section?" and "Which textual clue supports your interpretation?" Reception thus becomes an instructional approach that supports text-based, justified, and conscious meaning-making rather than restricting or controlling students' interpretations.

4. Generative Artificial Intelligence in Foreign Language Learning and Literary Reading

Generative AI can be considered a supportive tool in foreign language learning because it can provide explanations, offer feedback, rewrite texts, generate examples, and create individualized learning environments. Generative AI applications enable students to produce interactive and authentic content. In this respect, they go beyond the limits of traditional teaching methods and offer a more dynamic, participatory, and effective language learning process (Dinar, 2025). In this context, ChatGPT is regarded in foreign language classrooms not merely as a technical tool, but as a new learning partner that encourages the reconsideration of teaching and learning processes (Çınar Yağcı, 2026). Indeed, Slamet (2024) states that both teachers and students find ChatGPT useful for foreign language acquisition, motivation, and access to learning resources. Similarly, Mahapatra shows that ChatGPT can produce positive outcomes when used as a feedback tool in writing classes, but that effective use requires student guidance and instructional structuring. In the study by Özçelik and Çınar Yağcı (2024), ChatGPT's capacity to provide accurate and consistent feedback on grammatical errors in the development of French writing skills was examined. The study concluded that ChatGPT can be a useful support tool for identifying, correcting, and explaining many errors. However, it also emphasized that the tool should be used together with teacher guidance. Taken together, these findings suggest that generative AI is not an independent mechanism that replaces the teacher. Rather, when its purpose is clearly defined and its use is pedagogically supervised, it can function as a complementary tool that supports the learning process.

Generative AI, with functions such as explanation, feedback, and interactive question generation, may also support literary reading in foreign language learning. It can help students prepare for a text, understand unfamiliar vocabulary and context, compare perspectives, and develop initial interpretations. In this sense, it may reduce the difficulty students experience when first encountering a literary text. Mahapatra's emphasis on written feedback and structured learning, together with Slamet's focus on individualized study and immediate feedback, suggests that such tools can facilitate foreign language learning. However, this support should not replace

the student's own reading. Students should first encounter the text themselves, construct their own meaning, and then evaluate AI support by returning to the text. Literary reading is not the acceptance of a ready-made interpretation; it requires readers to engage with the text, follow textual clues, and justify their responses. Without reading journals and a text-based reader-response perspective, AI-generated interpretations cannot replace genuine reception (Carlisle, 2000, p. 12–13; Hirvela, 1996, p. 127–128). Thus, generative AI should be seen as a limited support tool that guides students back to the text and helps them justify their interpretations more consciously.

The most significant risk of generative AI in literary reading is that it may weaken students' direct and in-depth engagement with the text. Literary reading requires not only understanding what a text says, but also recognizing ambiguity, plurality of meaning, and interpretive possibilities. Zhai and colleagues' (2024) systematic review shows that unverified AI outputs may weaken critical thinking, analytical reasoning, and independent decision-making, while also raising risks related to misinformation, bias, plagiarism, privacy, and transparency. Similarly, Yalçın and Çınar Yağcı (2024) found that Bing Chat AI provided some accurate information about French culture, but that its responses were sometimes incomplete and insufficiently detailed; therefore, it could only partially support the development of French cultural competence when used with teacher guidance. Slamet (2024) also notes limitations concerning contextual understanding, accuracy, and technological dependence. These risks are particularly relevant to literary reading. If generative AI offers ready-made explanations before students form their own interpretations, it may reduce the ambiguity central to the literary text and weaken students' responsibility to verify meaning through the text. Thus, generative AI should not be treated as a final authority in literary interpretation, but as a second voice that students must question, compare with the text, and criticize when necessary.

The effective use of generative AI in literary reading depends largely on teacher guidance. Although these tools can support students' reading processes, unclear purposes, timing, or procedures may lead to superficial or inaccurate use rather than meaningful learning. Mahapatra (2024) notes that without proper guidance, the impact of AI tools remains limited and

the risk of misuse increases. Similarly, Slamet (2024) identifies teacher training, clear instructions, and continuous support as essential conditions for effective use. These findings show that generative AI should not be treated as an unrestricted technical tool, but as part of a teacher-planned reading process. The teacher should clarify which questions students may ask, which information they must derive directly from the text, and how they should evaluate AI outputs. Within such a framework, the tool can support vocabulary explanation, question diversification, contextualization, and comparison of interpretations. However, responsibility for understanding, interpreting, and justifying meaning must remain with the student.

In conclusion, generative AI may be useful in foreign language learning through functions such as explanation, feedback, and question generation. Yet in literary reading, what matters is not how many answers the tool provides, but how effectively it directs students back to the text. Literary reading is not limited to receiving ready-made answers. It requires students to think about the text, evaluate words in context, and explain their interpretations on the basis of textual evidence. If AI replaces the text, removes ambiguity, and turns interpretation into a ready-made conclusion, it narrows the student's space for thinking. By contrast, when used within a gradual reading process planned by the teacher to explain vocabulary, diversify questions, clarify context, and compare interpretations, it can support literary reception. For this reason, generative AI should not be used freely and without limits in literary reading. It should be integrated into a clear pedagogical framework that preserves students' direct engagement with the text.

5. Rationale for a Generative AI-Mediated Framework for Literary Text Reception

In foreign language classes, literary texts are often used in a limited way, mainly for vocabulary teaching, grammar practice, and comprehension questions. Such use may restrict students' ability to establish a deeper, more personal, and more critical relationship with the text. The traditional question-and-answer approach can turn literary reading into the search for a single correct answer and narrow students' interpretive space (Ünal, 2005, p. 205–207). Yet the mere presence of literary texts in the classroom is not

sufficient. What matters is why the text is selected, how it is approached, and within what kind of instructional structure it is used (Bulut, 2017; Ersan, 2019). Studies on French coursebooks support this point. Demirkan notes that cultural and artistic elements are included in coursebooks (2020, p. 541–545), while Kuşçu identifies activity spaces that allow students to engage in creative production (2020, p. 638–639). However, these elements do not automatically lead students to interpret, question, or connect the text with their own reading experience. Therefore, the main problem is not the absence of literary texts, but the limited pedagogical relationship often established with them.

A new instructional framework is needed to transform literary reading into a structured and meaningful learning process. Students should first be prepared for the text, then observe its linguistic, semantic, cultural, and discursive features, and finally justify their interpretations through textual evidence. Reading journal practices and reader-response-based tasks suggest that literary reading can be organized around preparation, encounter with the text, and justification of interpretation (Carlisle, 2000, p. 12–13; Garzón & Castañeda, 2015, p. 187–188). In French language teaching, this process should also include cultural and discursive evaluation, since literary texts produce meaning not only through language, but also through cultural signs, aesthetic choices, and social references. The poems, culture pages, and creative production activities found in French coursebooks show that this dimension is already present in instructional materials (Demirkan, 2020, pp. 549–553; Kuşçu, 2020, p. 638–639). However, students should not be expected to interpret the text immediately. They need to observe, infer, compare, and gradually support their interpretations with clear reasons. In this way, literary reading can develop linguistic awareness, cultural sensitivity, and critical thinking together.

A generative AI-supported framework should be integrated into this staged reading process rather than used as an independent or uncontrolled element. In the pre-reading stage, generative AI may help students understand key vocabulary, clarify brief historical and cultural background, and generate questions for the first reading. During reading, it can restate difficult structures more clearly or help students compare alternative interpretations of the same passage. Çınar Yağcı and Genç (2024) show that generative AI

tools can support not only explanation and feedback, but also production-oriented processes such as writing, translation, and visualization in French language teaching. Their study indicates that tools such as ChatGPT, DeepL, and Microsoft Designer may contribute to students' language skills, creativity, cultural awareness, and digital literacy in digital storytelling. In the post-reading stage, however, the central task is not to accept AI output as a ready-made answer, but to compare it with the student's own interpretation and with the literary text itself. For this reason, AI output should be treated as a provisional response to be examined, revised, or rejected when necessary. Given the need for teacher guidance, student preparation, conscious use, and the risk of overdependence, these tools should be used only in tasks whose aims, limits, and procedures are explicit and pedagogically controllable (Mahapatra, 2024; Zhai et al., 2024).

In literary reading, an interpretation becomes persuasive only when it is grounded in the text. Students should relate their claims to word choices, narrator perspective, images, repetitions, silences, or contextual clues. In this way, interpretation moves beyond personal impression and becomes a reasoned, critical, and text-based evaluation. Garzón and Castañeda's (2015, p. 187–188) emphasis on meaning-making and higher-order thinking, together with Bloemert et al.'s (2019, p. 378–381) finding that students view literature classes as a space for developing critical thinking, shows that the value of literary interpretation lies not only in originality, but also in textual grounding. At this point, generative AI should not interpret on behalf of the student. Rather, it should serve as a support resource that students question, compare, and, when necessary, criticize. Students should be able to explain why an AI-generated interpretation is supported or not supported by the text.

The need for such a framework does not arise from the desire to add a new technological tool to teaching. It arises from the need to balance two risks: reducing the literary text to vocabulary and grammar instruction, and transferring the interpretive process to AI. The first risk obscures the aesthetic, cultural, and interpretive layers of the text; the second weakens students' direct engagement with meaning and textual evidence. Generative AI should therefore be positioned not as a tool that thinks for the student, but as a limited support that redirects the student to the text. A balanced

instructional structure should preserve student participation, make teacher guidance visible, and limit AI to functions such as explanation, question generation, comparison, and feedback. Only within such a structure can literary texts in French language teaching become multidimensional learning spaces that develop interpretation, inference, and critical reading skills.

6. A Six-Stage Generative Artificial Intelligence-Supported Framework for the Reception of Literary Texts

In teaching French as a foreign language, the reception of literary texts is not a linear activity limited to comprehension. It requires students to prepare for the text, recognize their initial responses, analyze linguistic and structural features, interpret implicit meanings, evaluate different perspectives, and produce their own meaning. Therefore, a generative AI-supported framework should not reduce literary reading to a single task. It should gradually deepen the student's relationship with the text. The six-stage generative AI-supported framework (Figure 1) proposed here aims not to replace literary interpretation with generative AI, but to help students approach texts more consciously, attentively, and critically.

This framework distinguishes three roles. The student is the main subject who constructs meaning and develops interpretations through textual evidence. The teacher structures the reading process, deepens students' interpretations, and keeps the use of generative AI within pedagogical limits. Generative AI serves as a limited support tool for explanation, question generation, contextualization, alternative perspectives, and feedback. This distinction matters because literary reading is not the acquisition of ready-made knowledge, but an intellectual engagement with the text. Iser's (1972) emphasis on the reader's role in filling textual gaps and Hirvela's (1996) view that personal response must be justified through the text form the theoretical basis of this framework. Thus, generative AI has instructional value only when it redirects students back to the text.

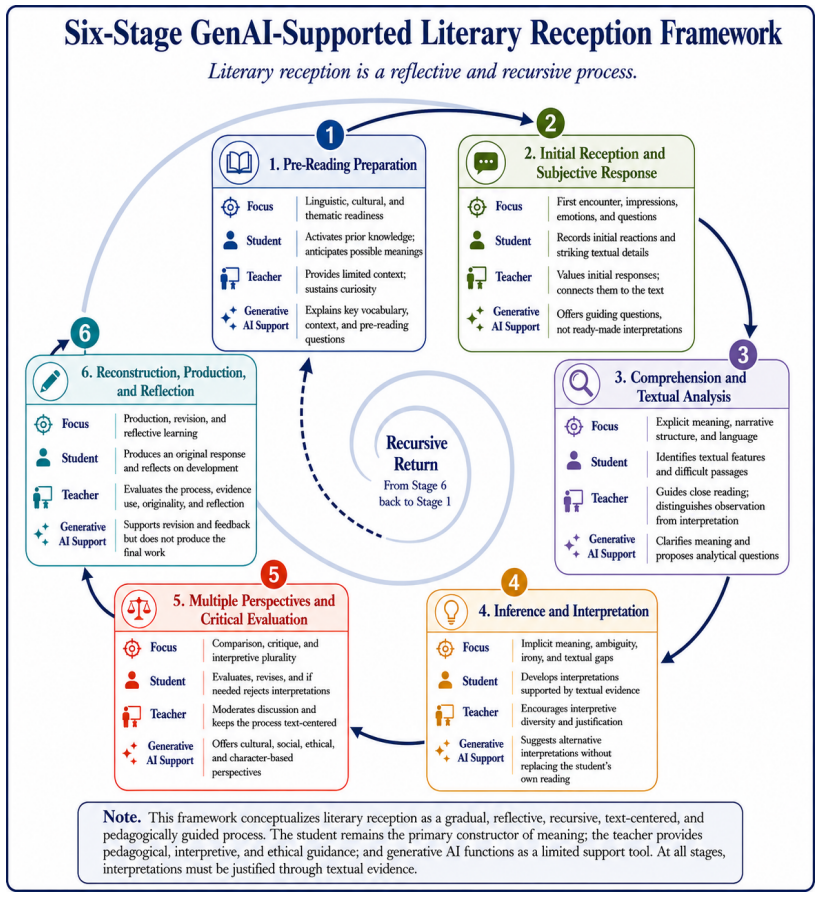


Figure 1 A Six-Stage Generative Artificial Intelligence-Supported Framework for the Reception of Literary Texts (Note. This figure was created with the assistance of ChatGPT-5.5 and subsequently reviewed and revised by the author.)

6.1. Pre-Reading Preparation

The first stage prepares students linguistically, culturally, and thematically before they encounter the text. The aim is not to provide all possible interpretations in advance, but to equip students with the minimum contextual knowledge, key vocabulary, and initial questions needed for reading. This is especially important for learners of French, since literary texts may include metaphors, cultural references, historical traces, genre-specific forms, and stylistic features that go beyond everyday language. Without

preparation, students may perceive the text merely as difficult linguistic material, which may limit reception from the outset.

At this stage, generative AI can facilitate entry into the text by explaining key concepts, giving brief contextual information, introducing cultural elements, or generating pre-reading questions. However, it should not interpret the text in advance. Its role is not to answer "What does the text mean?" but to support the question "With what prior knowledge can the student read this text more attentively?" Students should compare these explanations with their own knowledge, identify what they already know or find difficult, and write a short prediction based on the title, key words, or context. The teacher supervises this process, prevents premature interpretation, and preserves students' curiosity, since the value of the literary text lies not only in being explained, but also in generating questions.

6.2. Initial Reception and Subjective Response

The second stage focuses on the student's first encounter with the text and the subjective response it produces. Literary reading often begins with an impression created by a word, image, character, narrative voice, or ambiguous feeling. This response may be intuitive or fragmented, but it is often the first ground of interpretation. From a reader-response perspective, the student's personal response is the starting point of a deeper relationship with the text (Carlisle, 2000; Garzón & Castañeda, 2015).

At this stage, generative AI should not produce the first interpretation on behalf of the student. If the student asks the tool to interpret the text before reading it, the reception process is weakened from the beginning. Instead, the tool should only generate questions that help students articulate their initial responses, such as which expression surprised them, what tone they perceived, or where they felt the need to pause and rethink. The student's task is to express the first impression clearly, while understanding that it is not the final interpretation. The teacher should not immediately judge this response as right or wrong, but should connect it to the text through questions such as "Which expression created this feeling?" or "Which detail supports your first thought about this character?" In this way, subjective response becomes a starting point for text-based interpretation.

6.3. Comprehension and Textual Analysis

The third stage focuses on the explicit meaning and structural features of the text. Students examine what the text says, how events develop, how characters are positioned, from which perspective the narrator speaks, and how language is used. This stage provides the basis for interpretation, since interpretations made without understanding the text's explicit meaning and structure often become detached from it. The plurality of meaning in a literary text does not mean that it can be interpreted without limits; every interpretation should be connected to linguistic and structural signs.

Generative AI can support comprehension by explaining difficult words, rephrasing complex sentences, summarizing character relationships, or suggesting guiding questions about narration. Yet it should not complete the analysis for the student. Instead of producing a conclusion such as "the narrator is unreliable," the tool should help students identify the textual clues that might support such an inference. Prompts should therefore be process-oriented rather than result-oriented. Students should ask for questions, tasks, or analytical guidance, not ready-made interpretations. The teacher's role is to help students distinguish observation from interpretation: first, what is present in the text must be identified; then, these observations can be connected to possible meanings.

6.4. Inference and Interpretation

The fourth stage moves from explicit meaning to implicit meaning. Literary texts often invite readers to think through images, symbols, silences, gaps, irony, contrasts, and ambiguity. Iser's (1972) approach to textual gaps supports this stage, since readers complete meanings that are not directly stated by following textual clues. For this reason, inference and interpretation are central to literary reception.

Generative AI may be used here to make alternative interpretations visible, but only after students have first produced their own interpretation. If the tool offers ready-made meanings too early, it may weaken students' thinking process. Students should first write their own inference, then ask the tool for an alternative reading, a contrasting view, or questions that challenge their interpretation. In this way, AI becomes a second voice through which interpretation is tested, not the source of interpretation itself.

Students must support claims such as “the character is lonely” or “the text contains social critique” with words, images, events, or narrative choices in the text. The teacher should allow interpretive diversity, but should not treat every interpretation as equally valid. The value of interpretation depends on how strongly it is justified through textual evidence.

6.5. Multiple Perspectives and Critical Evaluation

The fifth stage requires students to evaluate the text from different perspectives. Literary texts are related not only to individual emotions and events, but also to social relations, cultural values, historical contexts, identity issues, ethical conflicts, and ideological positions. Therefore, students should not close the text with a single interpretation; they should compare different reading possibilities. This stage is central to critical reading.

Generative AI can help produce alternative cultural, social, ethical, or character-centered readings. However, these explanations should not be accepted directly. Students should compare each perspective with the text and decide which interpretation is stronger or weaker. Zhai and colleagues’ (2024) warning about the risks of excessive trust in generative AI is especially relevant here, because critical evaluation requires not only seeing different interpretations, but also questioning them. Students should be able to say, for example, that the tool reads the text through cultural conflict, whereas certain expressions foreground individual alienation more strongly. The teacher organizes discussion, brings different interpretations together, and prevents the discussion from drifting away from the text. This approach helps students view generative AI not as an authority, but as a debatable interpretive source, thereby strengthening both literary reading and AI literacy.

6.6. Reconstruction, Production, and Reflection

The sixth and final stage asks students to produce new work based on the reading process and to evaluate their own learning. Literary reception is not completed by understanding the text alone. It becomes visible when students produce a new statement, interpretation, text, or idea. This production may be written, oral, visual, or multimodal. Students may write an interpretive paragraph, create a short text from a character’s perspective, design

an alternative ending, relate the text to a contemporary context, or prepare a reflective reading journal.

Generative AI can support writing and revision by giving linguistic feedback, suggesting clearer phrasing, asking questions about coherence, or offering options for improvement. Yet production must remain the student's responsibility. The tool should not write the text for the student; it should help the student review and improve their own work. Students should therefore explain which suggestions they accepted or rejected and why. Reflection is also essential. Students should evaluate what they initially thought about the text, which details they noticed during analysis, how their interpretation changed, and where AI support was useful or limited. The teacher should assess both the final product and the process, including text-based evidence, interpretive coherence, critical evaluation, linguistic expression, original contribution, and reflective awareness.

In conclusion, the six-stage generative AI-supported framework enables literary texts to be used more consciously and deeply in teaching French as a foreign language. It is not based on technology use that distances students from the text, but on a pedagogical structure that continually redirects them back to it. Pre-reading preparation facilitates entry into the text; initial reception makes subjective response visible; comprehension and textual analysis provide a basis for interpretation; inference and interpretation reveal implicit meanings; multiple perspectives and critical evaluation encourage questioning of both the text and AI outputs; and reconstruction, production, and reflection transform reception into the student's own discourse. Thus, generative AI does not occupy the center of literary reading. Under teacher guidance, it becomes a limited and pedagogically meaningful support tool for the student's meaning-making process.

7. Roles of the Teacher, Student, and Generative Artificial Intelligence

The effective implementation of the proposed six-stage framework depends not only on the sequence of stages, but also on the clear definition of the roles of the student, the teacher, and generative AI. The reception of literary texts is interpretive and multilayered. If the student becomes passive,

meaning-making is reduced to AI output. If the teacher is limited to technical guidance, pedagogical depth is lost. If generative AI becomes the authority of interpretation, the processes of thinking, questioning, and justification are weakened. Therefore, the central principle of the framework is clear: the student constructs meaning, the teacher structures the process, and generative AI supports this process in a limited and controllable way.

7.1. The Role of the Student

In this framework, the student is not a receiver who merely understands the text. The student is the central subject who engages with the text, makes inferences, develops interpretations, evaluates different perspectives, and produces a response. This role reflects the nature of literary reading. A literary text is not a closed structure whose meaning is already complete. As Iser (1972) emphasizes, textual gaps, ambiguities, and implied meanings are completed through the reader's participation. Therefore, students reading a literary text in a foreign language do not merely decode words; they also make sense of associations, silences, cultural references, and narrative organization.

The student's first responsibility is to establish direct contact with the text. GenerativeAImay support preparation or post-reading work, but it should not replace the first encounter with the text. Students should first notice their own questions, intuitions, difficulties, and initial responses. Carlisle's (2000) reading journal approach is useful here because it makes visible what students think, where they struggle, and which expressions affect them during reading. The second responsibility is to ground interpretation in the text. Students should connect their claims to words, images, repetitions, narrator choices, character behavior, or cultural signs. As Hirvela (1996) notes, personal response matters, but without textual grounding it may remain superficial. The third responsibility is to evaluate AI output critically. Students should compare it with the text, identify its strengths and weaknesses, and challenge it when necessary. In this way, they read both the literary text and the secondary text produced by AI, developing literary reading skills and AI literacy.

7.2. The Role of the Teacher

The teacher is the pedagogical center of this framework, but not the person who transmits the single correct interpretation. The teacher guides the student's relationship with the text, deepens interpretation, and defines the limits of generative AI use. In teaching literary texts, the teacher's task is not to lead students toward a predetermined conclusion, but to help them think on the basis of the text. Ünal's (2005) criticism of directing students toward a single correct interpretation is therefore central, since such an approach restricts creative and critical participation.

The teacher's first task is to structure the reading process in stages. Students should neither be left unprepared before reading nor directed by ready-made explanations before encountering the text. The teacher should clearly organize pre-reading preparation, initial reception, textual analysis, inference, critical evaluation, and production. Students need to know when they should write their own interpretation, when they may use AI, and how they should question its output. The second task is to supervise AI use pedagogically. Mahapatra (2024) and Slamet (2024) show that teacher guidance and clear instructions are necessary for effective use. For this reason, process-oriented prompts should be preferred over prompts that produce ready-made interpretations. The third task is to protect the quality of interpretation. Literary reading is open to interpretation, but not every interpretation is equally strong. Its value depends on textual evidence, coherence, and relation to the linguistic and cultural structure of the text. The teacher should also explain ethical issues, including the limits of AI use, citation, original production, and academic integrity.

7.3. The Role of Generative Artificial Intelligence

In this framework, generative AI is not an independent teacher, a definitive interpretive authority, or a subject that thinks on behalf of the student. Its role is limited to supporting the relationship students establish with the literary text. It can explain, generate questions, provide context, suggest alternative perspectives, offer linguistic feedback, and help students reconsider their interpretations. However, it should never determine the final meaning of the text.

This positioning is necessary because of the limitations of generative AI. As Zhai and colleagues (2024) state, using AI outputs without critical evaluation may weaken analytical thinking and independent decision-making. In literary reading, this risk is particularly important because literary texts are often based on ambiguity, plurality, and openness to interpretation. When AI turns these ambiguities into quick and definitive explanations, it may reduce the intellectual tension of the text. Therefore, AI output should be treated as a provisional suggestion open to discussion, not as a conclusion. Its most appropriate function is to redirect students back to the text. If the tool helps students look more closely at an expression, support an interpretation with evidence, or evaluate an alternative perspective, it is pedagogically useful. If students rely only on AI summaries without reading the text, literary reception is weakened. For this reason, generative AI is not the center of literary reading; it is a secondary tool that supports a process whose center remains the student, the text, and teacher guidance.

8. Practical Implications for French Language Teaching

The proposed framework is not only theoretical; it also offers an applicable reading structure for classes in which French is taught as a foreign language. Its main aim is to move the literary text beyond comprehension questions and to help students establish first a personal, then an analytical, interpretive, and critical relationship with the text. Generative AI does not need to be used at every stage. It should be included only when it is pedagogically meaningful. The central criterion is therefore whether AI distances students from the text or helps them return to it more attentively.

8.1. Designing Artificial Intelligence-Supported Literary Reading Activities

The first step in designing a French lesson based on a literary text is text selection. The text should be appropriate for students' language level, age, interests, and cultural readiness. If it is too long, dense, or culturally opaque, students may struggle to participate in the reception process. If it is overly simplified, however, its literary value and interpretive potential may weaken. The teacher should therefore balance linguistic accessibility with interpretive richness.

The second step is to plan the lesson in stages. Before reading, students may focus on key words, basic context, and their expectations about the text. Generative AI can support this stage by producing brief cultural explanations, vocabulary tasks, or prediction questions. During reading, students identify the text's explicit meaning, narrator, characters, time, setting, and notable linguistic features. AI may clarify complex expressions or generate textual analysis questions. After reading, students develop an interpretation, support it with textual evidence, and, when appropriate, compare it with alternative interpretations offered by AI.

The third step is to provide clear guidelines for AI use. Students should know when they may use the tool, which prompts are appropriate, and how to evaluate the output. A prompt such as "Interpret the text for me" should be avoided. Instead, students may use prompts such as "Which textual evidence can strengthen my interpretation?" or "What questions can be asked to examine the narrator in this text?" In this way, generative AI becomes not a source of ready-made answers, but a support mechanism that organizes student thinking.

8.2. Sample In-Class Activities

This framework can be adapted to different French levels. At A1–A2 levels, short poems, dialogues, micro-stories, or brief cultural texts may be used. At B1–B2 levels, short stories, novel excerpts, drama passages, or texts from Francophone literature may be selected. Activities can be simplified according to level, but the interpretive and reception-oriented dimension should be preserved.

In a pre-reading prediction activity, the teacher gives students the title and several key words. Students first write brief predictions about the topic, emotional tone, and possible conflict. Then generative AI produces three pre-reading questions based on the same title and key words. Students compare these questions with their own predictions. The purpose is not to explain the text, but to increase curiosity and expectation.

In a vocabulary and context activity, students identify difficult words in the text. AI may be asked to provide level-appropriate explanations and example sentences rather than direct translations. Thus, vocabulary work

remains connected to the text, and students begin to consider the function of words in context.

In an interpretation card activity, students write a claim about the text, such as “The narrator approaches the character from a distance” or “The feeling of loneliness is foregrounded.” They then find two pieces of textual evidence. AI is asked to generate an alternative question that challenges the claim, and students reconsider their interpretation. This activity helps students both construct and test interpretation.

In a comparative interpretation activity, students first write their own interpretation, then ask AI to generate a different reading of the same text. They compare the two interpretations in terms of textual evidence, coherence, and depth, and justify which is stronger. This prevents passive acceptance of AI output.

In a reflective writing activity, students write a short evaluation at the end of the lesson. They explain how their first impression changed, which textual details shaped their interpretation, and where AI support was useful or insufficient. Thus, students evaluate both the text and their own reading process.

8.3. Assessment and Reflection

Assessment in this framework should not be limited to correct answers. Since the reception of literary texts is not the search for a single fixed meaning, the quality of students’ engagement with the text, the development of their interpretation, and the evidence they use should be evaluated. Criteria may include comprehension, use of textual evidence, interpretive coherence, critical comparison, linguistic expression, original contribution, and reflective awareness.

Process assessment is especially important in AI-supported activities. Students may be asked to submit reading notes, an initial interpretation draft, the AI output they used, their critique of that output, and the final version of their text. This allows the teacher to follow the student’s thinking process and helps students recognize their own development. It also supports academic integrity, since students must explain when and why they used AI.

9. Ethical and Critical Dimensions

The use of generative AI in the reception of literary texts is both pedagogical and ethical. These tools may affect how students access information, interpret texts, produce written work, and assume responsibility for learning. Therefore, the aim should not be to reduce students' cognitive effort, but to make their thinking process more conscious and visible. Otherwise, generative AI may become a shortcut that replaces literary reading.

The first ethical issue is preserving students' responsibility to think. Literary reading requires students to engage with the text, tolerate ambiguity, test their own interpretations, and justify them. Ready-made explanations from AI may shorten this process, but this does not necessarily mean learning. If students use AI before their own reading, the reception process may become superficial. Therefore, the teacher should position AI use after the first reading and initial interpretation attempt.

The second issue concerns accuracy and reliability. Generative AI can produce persuasive language, but its outputs are not always accurate, sufficient, or contextually appropriate. In literary texts, historical context, cultural references, aesthetic orientation, and textual ambiguity may be simplified or misrepresented. Students should therefore compare AI output with the text and reliable sources. Yalçın and Çınar Yağcı's (2024) study on French cultural competence found that AI tools may provide some accurate information, but that this information can sometimes remain incomplete and insufficient in detail. This finding shows why teacher guidance is indispensable in literary and cultural content.

The third issue is originality and academic integrity. If students present an interpretation produced by AI as their own, both learning and fair assessment are damaged. Clear classroom principles are therefore necessary. Students should explain at which stage they used AI, what output they received, and how they evaluated it. This declaration is not only an ethical requirement, but also a pedagogical tool that makes thinking visible.

The fourth issue is data privacy and text sharing. Students should not upload personal information, classmates' data, or long texts that may raise copyright concerns. The teacher should clearly explain what should not be

shared. In educational uses of digital tools, privacy, data security, and terms of use must always be considered.

Teacher guidance is therefore the key condition for ethical use. If AI-supported literary reading is left unrestricted, it may create problems for academic integrity and learning quality. When used with clear guidelines, staged tasks, and reflective assessment, however, generative AI can support both literary reading and digital literacy.

10. Discussion

The six-stage generative AI-supported framework proposed in this study contributes to the teaching of literary texts in French as a foreign language by rethinking literary reception within contemporary digital learning conditions. Its main contribution lies in moving beyond two reductive positions. The first reduces literary texts to vocabulary, grammar, and comprehension exercises, although literature in language education offers broader linguistic, cultural, discursive, and interpretive possibilities (Bulut, 2017; Hall, 2015; Bloemert et al., 2019). The second treats generative AI as a source of immediate interpretations and ready-made explanations, a use that may weaken students' critical thinking and independent judgment when outputs are accepted without careful evaluation (Zhai et al., 2024). Against these two positions, the framework considers the literary text as a multilayered space of linguistic, cultural, aesthetic, and interpretive engagement, while positioning generative AI as a controlled support that redirects students to the text. In this respect, the model differs from traditional uses of literature in foreign language teaching. It does not approach the literary text as an object to be explained by the teacher or decoded through predetermined questions. Nor does it leave interpretation entirely to subjective response. Instead, it organizes reading as a gradual process in which initial impressions become visible and are then transformed into text-based, justified, and critical interpretations.

This point is important because literary reception requires both openness and discipline. From the perspective of reception theory, the literary text is not a closed structure that contains a single fixed meaning; rather, meaning emerges through the reader's expectations, responses, and participation in filling textual gaps (Iser, 1972). In foreign language teaching, this view

supports a more active understanding of reading, since students are not expected merely to find a predetermined meaning, but to construct meaning through their encounter with the text (Ünal, 2005). However, personal response alone is not sufficient. Reader response becomes pedagogically meaningful when it is connected to textual evidence and transformed into critical reflection (Hirvela, 1996). Similarly, reading journals and reader-response-based activities show that students' first impressions, questions, hesitations, and interpretations can become visible and open to further development when they are pedagogically structured (Carlisle, 2000; Garzón & Castañeda-Peña, 2015). Therefore, students should be allowed to produce personal meanings, but these meanings must be supported through words, images, narrative choices, silences, cultural signs, and other textual evidence. The framework thus extends reader-response pedagogy by linking subjective response, textual analysis, critical evaluation, and reflective production within a coherent instructional sequence.

The integration of generative AI also requires a reconsideration of reception theory in foreign language learning. In classical reception approaches, meaning emerges through the interaction between the reader and the text (Iser, 1972). In an AI-supported environment, this interaction gains an additional layer because the student also encounters a secondary discourse produced by AI about the literary text. This secondary discourse does not replace reception, but it changes its conditions. Literary reading becomes a process in which students interpret the text, examine AI output, compare both, and decide which interpretation is more coherent and better supported. Thus, AI-supported reception should not be understood as automated interpretation, but as a form of mediated reception that requires critical comparison and epistemic vigilance. This point also clarifies why teacher guidance in the framework is not merely technical. The teacher does not only decide when students may use AI or which prompts they may write. More importantly, the teacher defines the conditions under which an interpretation can be considered valid, limited, weak, or persuasive. Previous studies similarly underline that AI tools become pedagogically useful only when their use is guided, structured, and supported by clear instructions (Mahapatra, 2024; Slamet, 2024). In this sense, teacher guidance has an epistemological function. It helps students understand how literary

meaning is constructed, how evidence is selected, how interpretations are justified, and why AI output cannot be accepted as an unquestioned authority.

The framework also redefines the student's interpretive responsibility in the age of generative AI. In conventional literary reading, the student is expected to understand the text and develop an interpretation. In an AI-supported setting, this responsibility becomes more complex. The student must first encounter the text directly, formulate an initial response, analyze textual features, and construct an interpretation. Only after this process may AI be used to ask questions, compare perspectives, receive linguistic support, or test the coherence of an idea. Such uses are consistent with studies showing that generative AI can support explanation, feedback, individualized learning, writing, translation, and production-oriented language tasks when used within pedagogical limits (Çınar Yağcı & Genç, 2024; Mahapatra, 2024; Özçelik & Çınar Yağcı, 2024; Slamet, 2024). However, the final responsibility for meaning remains with the student. The student must explain why an AI-generated suggestion is accepted, rejected, revised, or only partially used. In this way, literary competence and AI literacy develop together. The proposed framework is therefore significant because it offers a balanced position between resistance to technology and uncritical technological adoption. Excluding generative AI entirely from literary reading may ignore contemporary learning realities, whereas placing it at the center of interpretation may weaken students' direct engagement with the literary text and increase the risk of overdependence, misinformation, or superficial reasoning (Zhai et al., 2024). The framework addresses this tension by limiting AI to specific pedagogical functions: preparation, explanation, question generation, comparison of perspectives, linguistic feedback, and reflective support. These functions are valuable only when they deepen students' reading process; they become problematic when they replace students' own interpretive effort.

Another implication of the framework concerns assessment. If literary reception is viewed as a process, evaluation should not be limited to the final written product or the correctness of an answer. It should also consider how students move from first impressions to interpretation, use textual evidence, assess alternative readings, and reflect on AI support. In this sense, the model promotes process-based assessment through reading notes, initial

interpretations, AI outputs, students' critiques of these outputs, and revised responses. This approach is consistent with reading journal practices that make interpretive development visible (Carlisle, 2000) and supports academic integrity by making AI use transparent and accountable. Nevertheless, the framework remains a conceptual and pedagogical proposal rather than a validated instructional model. Its effectiveness may vary according to language level, literary genre, institutional context, teacher competence, and access to digital tools. At beginner levels, AI support may focus on vocabulary, context, and simple guiding questions, while at advanced levels it may support comparative interpretation, critical discussion, and reflective writing. Future classroom-based studies should examine its effects on interpretation depth, textual evidence use, motivation, writing, critical thinking, and students' ability to evaluate AI output. Teacher experiences should also be explored, since implementation depends on planning, task design, and careful moderation. Overall, the framework shows that generative AI can support literary reading without reducing literature to information retrieval or interpretation to automated output. Its central contribution is to frame AI-supported literary reception as a pedagogically mediated process in which students learn to question, compare, justify, revise, and critically control meaning.

11. Conclusion

This study proposed a generative AI-supported instructional framework for the reception of literary texts in teaching French as a foreign language. The main argument is that literary texts should not be treated merely as auxiliary materials for vocabulary and grammar instruction. Rather, they should be regarded as rich learning spaces through which students encounter language in context, recognize cultural signs, make inferences, develop perspectives, and support their interpretations with textual evidence. However, the presence of literary texts in the classroom does not by itself ensure meaningful learning. Text selection, question design, student participation, and teacher guidance are decisive. If literary reading is reduced to a traditional question-and-answer structure, the text may become closed content with a single expected meaning, and students' personal, intellectual, and critical engagement may weaken.

The proposed six-stage framework aims to prevent this limitation by organizing literary reading as a staged, explicit, and participatory process. Pre-reading preparation, initial reception, comprehension and textual analysis, inference and interpretation, multiple perspectives and critical evaluation, and reconstruction, production, and reflection gradually deepen students' relationship with the text. Within this process, generative AI may support vocabulary explanation, contextualization, question generation, alternative perspectives, and feedback. Yet it should not read or interpret on behalf of the student. Its outputs must be compared with the literary text, questioned, and evaluated under teacher guidance. Thus, the framework preserves student agency, strengthens literary reception, and clarifies the pedagogical boundaries of AI use in foreign language teaching.

12. Limitations and Future Research

This study proposes a theoretical and instructional framework for the reception of literary texts in teaching French as a foreign language. Its main limitation is that the model has not yet been tested through classroom-based implementation. It was developed on the basis of relevant literature, reception theory, foreign language teaching, and current discussions on generative AI. Therefore, empirical studies are needed to examine how the framework functions with different student groups, proficiency levels, and literary genres. Future research may first implement the model at A1–A2 levels with short, accessible, and culturally supported texts, and at B1–B2 levels with denser narratives, poems, drama passages, or Francophone literary texts. Such studies would help determine which AI functions are more appropriate and effective at each level.

Further research should also examine both student and teacher experiences. Studies may investigate how students perceive AI-supported literary reading and how this process affects their motivation, confidence in interpretation, use of textual evidence, written production, and tendency to return to the text. Interviews, reflective journals, student products, and classroom observations may provide useful data. Teacher experiences are equally important, since the success of the framework depends on text selection, task design, limits placed on AI use, and assessment practices. In addition, comparative and longitudinal studies may explore differences between AI-

supported and non-supported literary reading activities, as well as students' development of more independent, critical, and text-based interpretations over time. These future studies can help transform the proposed model from a conceptual framework into an adaptable and classroom-based instructional approach.

Declaration on the Use of Generative Artificial Intelligence

ChatGPT (OpenAI, GPT-5.5) was used solely to support translation, linguistic editing, and figure creation during the preparation of this chapter. The theoretical framework, scientific content, source selection, and final interpretations were developed, reviewed, and revised by the author. The author takes full scientific and ethical responsibility for the chapter.

References

Avrupa Konseyi. (2021). *Diller için Avrupa Ortak Başvuru Metni: Öğrenme, öğretme ve değerlendirme: Tamamlayıcı cilt* (S. Çelik, Y. Kırkgöz, D. Çopur, P. İrgin ve Ş. Erbay Çetinkaya, Çev.). Millî Eğitim Bakanlığı. (Orijinal eser 2020 yılında yayımlandı)

Bloemert, J., Paran, A., Jansen, E., & van de Grift, W. (2019). Students' perspective on the benefits of EFL literature education. *The Language Learning Journal, 47*(3), 371–384.

Bulut, T. (2017). Yabancı dilin öğretilmesinde edebiyatın önemi. *Aydın Türklük Bilgisi Dergisi, 3*(1), 1–9.

Carlisle, A. (2000). Reading logs: An application of reader-response theory in ELT. *ELT Journal, 54*(1), 12–19.

Çınar Yağcı, Ş. (2026). L'Intelligence Artificielle Generative Dans L'enseignement/Apprentissage Des Langues Etrangeres : Tendances, Principaux Resultats Et Perspectives De Recherche. In N. Özçelik (Ed.), Researches Internationales en Didactique du Français Langue Etrangere-I (pp. 49-95). Eğitim Yayınevi.

Çınar Yağcı, Ş., & Genç, A. (2024). Üretken yapay zekâ araçları ile Fransızca dijital hikâye tasarlama: Yazma, görselleştirme ve çeviri süreçleri. N. Özçelik (Ed.), *Fransız dili eğitimi alanında uluslararası araştırmalar I / International research in the field of French language teaching* içinde (ss. 85–110). Eğitim Yayınevi.

Demirkan, D. (2020). Yabancı dil Fransızca öğretimi ders kitabı "Le Nouveau Taxi 1"de A 1 düzeyi kültürel öğelerin sunumu. *RumeliDE Dil ve Edebiyat Araştırmaları Dergisi*, Ö7, 541–556.

Dinar, G. (2025). Yapay zekâ ve eylem odaklı yabancı dil öğretimi: Kuram, uygulama ve gelecek perspektifleri. Y. Topaloğlu (Ed.), *Yapay zekâ ve yabancı dil eğitimi: Kuramsal temeller, uygulamalar ve yeni yaklaşımlar* içinde (ss. 54–79). Günce Yayınları.

Ersan, Y. (2019). Yabancı dil öğretiminde kısa metinlerin kullanımı ve önemi. *Karabük Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 8*(2), 288–296.

Garzón, E., & Castañeda-Peña, H. (2015). Applying the reader-response theory to literary texts in EFL-pre-service teachers' initial education. *English Language Teaching, 8*(8), 187–198.

Hall, G. (2015). *Literature in language education* (2nd ed.). Palgrave Macmillan.

Hirvela, A. (1996). Reader-response theory and ELT. *ELT Journal, 50*(2), 127–134.

Iser, W. (1972). The reading process: A phenomenological approach. *New Literary History, 3*(2), 279–299.

Kuşçu, E. (2020). Yabancı dil olarak Fransızca öğretiminde sosyokültürel öğelerin öğrenme süreçlerine katkısı: "Francofolie 1" örneği. *RumeliDE Dil ve Edebiyat Araştırmaları Dergisi, 21*, 633–649.

Mahapatra, S. (2024). Impact of ChatGPT on ESL students' academic writing skills: A mixed methods intervention study. *Smart Learning Environments, 11*, Article 9.

Özçelik, N., & Çınar Yağcı, Ş. (2024). L'enseignement des langues étrangères soutenu par l'intelligence artificielle générative: L'effet de ChatGPT sur l'amélioration des compétences rédactionnelles en français. N. Özçelik (Ed.), *Fransız dili eğitimi alanında uluslararası araştırmalar I / International research in the field of French language teaching* içinde (ss. 25–60). Eğitim Yayınevi.

Slamet, J. (2024). Potential of ChatGPT as a digital language learning assistant: EFL teachers' and students' perceptions. *Discover Artificial Intelligence, 4*, Article 46.

Tilbe, A. (2019). Yeniötesi yazında özkurmaca. Transnational Press London.

Topaloğlu, Y. (2023). Yabancı dil olarak Fransızca yazılı üretimi bağlamında gelecekteki Türk öğretmenlerin sözdizimi, dilbilgisi ve yazım yanlışları. *Söylem Filoloji Dergisi, 8*(3), 1207–1225.

Tsang, A., Paran, A., & Lau, W. W. F. (2023). The language and non-language benefits of literature in foreign language education: An exploratory study of learners' views. *Language Teaching Research, 27*(5), 1120–1141.

Ünal, D. Ç. (2005). Yabancı dil öğretiminde edebi metinler: Yenilikçi yaklaşımlara geçiş süreci ve gerekçeleri. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 29*, 203–212.

Yalçın, P., & Çınar Yağcı, Ş. (2024). Fransız kültür edincinin kazandırılmasında yapay zekâ rehber olabilir mi? Bing Chat AI örneği. *Yükseköğretim ve Bilim Dergisi / Journal of Higher Education and Science, 14*(2), 308–327.

Zhai, C., Wibowo, S., & Li, L. D. (2024). The effects of over-reliance on AI dialogue systems on students' cognitive abilities: A systematic review. *Smart Learning Environments, 11*, Article 28.

Designing Classroom Activities and Games for Chatbot-Supported Task-Based Language Teaching

Aybegüm Aksak[1] & *Selçuk Doğan*[2]

1. Introduction

Rapid advances in digital technologies have transformed language education, as they have many other fields. Online learning environments have become increasingly widespread. Mobile technologies are now indispensable in everyday life. Artificial intelligence applications have also entered educational processes. These developments require methods, materials, and learning environments to be reconsidered and reorganised. Remote and hybrid learning models have now become established. Consequently, technology-enhanced practices are no longer merely alternatives. They are core components of language teaching (Şengül, 2021). Education can no longer be conceived independently of artificial intelligence and digital learning tools.

Research conducted in the context of digital transformation has demonstrated the contribution of digital tools to language teaching. Web 2.0 technologies, digital storytelling applications, online learning environments, interactive materials, and mobile applications enable learners to engage with the target language more intensively and continuously. While these tools have largely freed learning processes from temporal and spatial constraints, they have also facilitated learner responsibility and individual learning (Genç Ersoy & Ersoy, 2021). Nevertheless, online learning environments may also give rise to problems such as limited interaction, reduced

[1] Lecturer Dr., Ankara Yıldırım Beyazıt University, ORCID: 0000-0002-5367, *aybegumaksakgonul@aybu.edu.tr*, Türkiye

[2] Lecturer Dr., Yozgat Bozok University, ORCID: 0000-0003-2502-5776, *selcuk.dogan@bozok.edu.tr*, Türkiye

motivation, and declining communicative participation (Şengül, 2021). Technology use should therefore be approached not merely as a technical innovation, but in conjunction with pedagogical approaches designed to enhance learner engagement and interaction.

Generative artificial intelligence is among the most notable recent developments in educational technology. Chatbots based on large language models have created new opportunities in language education. They also have the potential to facilitate instruction. Kasneci et al. (2023) identify several significant educational opportunities offered by large language models. These include content generation, increased engagement, personalised learning, and support for feedback. ChatGPT, Gemini, Claude, and similar applications can conduct human-like interactions in natural language. These interactions can reflect real-life situations. The applications can also provide immediate feedback and generate content based on learner needs. They can therefore offer personalised learning experiences. Integrating chatbots for different pedagogical purposes is one effective use of artificial intelligence in language education.

AI-supported chatbots should not be used independently of pedagogical objectives. Instead, they should be integrated with established teaching and learning theories. Task-Based Language Teaching (TBLT) provides an important theoretical framework for such integration. TBLT conceptualises language learning as the performance of meaningful communicative tasks. It therefore extends beyond the study of grammar. Learners exchange information, make decisions, and use the target language for communication. They do so while solving likely everyday problems and completing relevant tasks. Chatbots can facilitate natural-language interaction during these tasks. Learners may interview a chatbot, devise a travel plan, or solve a problem. They may also develop a story or enact an assigned role. Such practices strengthen the communicative dimension of language teaching. Language use consequently becomes less artificial and constrained. It also assumes a more authentic and communication-oriented character.

Gamification is another area connecting task-based learning with AI-supported chatbots. It introduces game elements into non-game instructional environments. These elements include points, levels, tasks, narratives, and rewards. Their purpose is to increase learners' active engagement. Chatbots

can present such elements interactively and in personalised forms. Gamified applications can therefore generate level-appropriate tasks and performance-responsive scenarios. They can also provide immediate feedback. These features reduce monotony and enhance learner motivation. Learning then becomes an experience based on exploration, problem-solving, and communication.

AI-supported chatbots offer several potential contributions to teaching Turkish as a foreign language. They can support speaking practice and provide feedback on writing. They can also contribute to vocabulary instruction, intercultural interaction, and communicative competence. Nevertheless, research on chatbot use in this field remains limited. This increases the importance of relevant theoretical and practice-oriented studies.

2. Theoretical Framework

2.1. Core Principles of Task-Based Language Teaching

Task-Based Language Teaching (TBLT) emerged from communicative and action-oriented approaches. It is used in second and foreign language education. The approach centres on using language for meaningful purposes. Its primary aim is functional target-language use in everyday situations.

In TBLT, a task is a coherent piece of work with a communicative purpose. It differs from activities that merely practise predetermined grammatical structures. Ellis (2003) identifies meaning as the primary focus of a task. Learners use their own linguistic and non-linguistic resources. A task also contains a communicative or information gap requiring resolution. Moreover, it produces a clearly defined outcome beyond language use. Nunan (2004) similarly emphasises comprehension, production, and interaction in the target language. Learners focus on conveying meaning rather than practising particular forms. Long (2015) derives tasks from real-world needs beyond the classroom. Skehan (1998) identifies meaning, problem-solving, real-life relevance, and outcome-based evaluation as defining criteria. Willis and Willis (2007) likewise emphasise functional language use and a tangible outcome. Possible outcomes include a list, plan, story, decision, or problem solution. A task is therefore not defined simply by being enjoyable

or interactive. It combines meaning-oriented communication, information exchange, authentic language use, and an observable outcome.

Learner-centredness, a common feature of contemporary approaches, is also a core principle of TBLT. Learners assume responsibility for planning, performing, and evaluating tasks and are therefore active participants in the process. This enables them to exercise greater control over their own learning and enhances their motivation. A study conducted by Nget et al. (2020) with secondary-school learners in Thailand found the following. Task-based instruction increased classroom participation and led to significant improvement in speaking skills.

A further core principle of TBLT is that language learning takes place through meaningful tasks. The concept of a task refers to activities that require learners to use language for a specific communicative purpose. Such tasks may involve solving a problem, exchanging information, making a decision, creating a product, or communicating towards a particular goal. Their defining characteristic is that learners' attention is directed towards meaning and function rather than grammar. Tasks are therefore based on authentic situations that reflect real-life communication (Michel, 2023). Learners' linguistic development occurs as they use language as a natural tool for completing the assigned task.

Another important principle of task-based instruction is its focus on developing communicative competence. Communicative competence encompasses not only grammatical accuracy but also appropriate lexical choice, the ability to construct discourse, contextually appropriate language use, and communication strategies. Through group work, role-play, discussion, and problem-solving tasks, task-based language practice provides learners with opportunities to develop this competence. Lume and Hisbullah (2022) found the following. Task-based instruction produced more positive outcomes for learners' speaking skills than traditional methods. They note in particular that tasks based on the negotiation of meaning increased learners' fluency and communicative confidence.

Focus on form is another principle that occupies an important place in the theoretical foundations of TBLT. In this approach, attention is drawn to particular linguistic forms in response to the linguistic needs that arise. Learners are performing tasks. Grammar is thereby learnt not as a body of

abstract rules, but as a natural component of communicative need (Michel, 2023). This approach contributes to establishing a balance between linguistic accuracy and communicative fluency.

Task design is likewise central to the approach. Effective tasks should be meaningful, attainable, relevant to real life, and appropriate to learners' proficiency levels. Although tasks that learners find engaging can increase participation, excessively complex tasks may impose a high cognitive load. Task design must therefore take account of learners' language proficiency, needs, and learning objectives.

A further defining feature of TBLT is its support for collaborative learning. Tasks frequently require learners to work in pairs or small groups, during which they share information, make joint decisions, and produce an outcome together. Research on online collaborative writing, in particular, indicates the following. Task-based practice increases learners' language use and supports higher-order thinking skills (Wang & Lynch, 2022). This process of learning through social interaction is also consistent with the fundamental assumptions of sociocultural learning theory.

Technological developments have also had a substantial influence on how task-based instruction is implemented. Technology-mediated tasks have become increasingly prevalent in language education, particularly with the expansion of online learning environments. Digital platforms enable learners to communicate across different times and locations. While multimedia tools enrich task design and performance. Mulyadi et al. (2023) found the following. Technology-enhanced tasks developed learners' self-regulation skills and fostered more positive attitudes towards the learning process. These findings support current efforts to integrate AI-supported chatbots and digital learning environments with TBLT.

Performance-based assessment is another core principle of the task-based approach. Whereas traditional examinations often measure learners' knowledge of grammar, assessment in TBLT focuses on the extent to which learners can accomplish tasks. Performance tasks, scoring rubrics, portfolios, and project work are therefore frequently employed. Noroozi and Taheri (2022) state that task-based assessment practices are more effective than traditional assessment methods in measuring learners' real-world

language use. Assessment thus becomes not only a means of evaluating learning outcomes, but also an integral part of learning.

The proliferation of digital technologies and artificial intelligence applications has increased the potential of TBLT to provide more flexible, interactive, and personalised learning experiences. The task-based approach is therefore regarded as one of the most important theoretical foundations for chatbot-supported language education and gamified learning environments.

2.2. Pedagogical Agents and Chatbot Technologies in Language Education

Advances in artificial intelligence, natural language processing (NLP), and large language models (LLMs) have transformed language education. They have enabled new learning ecosystems to emerge. AI-supported tools can make learning more personalised and effective. They also extend traditional methods and respond to future educational needs (Çangal et al., 2025, p. 85). Pedagogical agents and chatbots occupy an important place in this transformation. They provide personalised support, interactive experiences, and opportunities for real-time language use. Pedagogical agents include digital characters, virtual tutors, and robotic systems. They guide learners, provide feedback, and support learning. Chatbots are AI-based systems that conduct human-like dialogue through natural language processing.

The use of chatbots in language education can provide learners with a continuously available environment for communication. Whereas opportunities to use the target language in conventional classrooms are often restricted to lesson time, chatbots can function as virtual conversation partners available at any time of day. In their systematic review of chatbot-supported language learning, Huang et al. (2022) report that chatbots offer important technological, pedagogical, and social benefits. They can provide immediate feedback and personalised learning experiences while enabling learners to communicate more extensively in the target language. Their capacity to create low-anxiety learning environments and support learner autonomy is regarded as one of their principal advantages in foreign language education.

The functions of pedagogical agents in language education are not confined to speaking practice. Contemporary AI-supported systems can identify learners' grammatical errors, evaluate written texts, provide pronunciation feedback, and generate content appropriate to learners' proficiency levels. These capabilities offer substantial advantages for personalised learning. Research on AI-supported language learning environments indicates the following. Learners can undertake activities at a pace suited to them and receive more intensive support in areas where it is needed. Instruction can thus move away from standardised practice towards a dynamic structure responsive to individual learner needs.

The emergence of large language models has considerably expanded the capabilities of chatbots used in education. Generative AI systems such as ChatGPT can engage learners in multi-turn dialogue, produce creative content, and offer feedback adapted to learner performance. They have also been reported to increase learner motivation, support active engagement in learning, and facilitate teachers' lesson preparation. The same research, however, draws attention to risks including the generation of inaccurate information, cultural bias, and over-reliance.

Another important contribution of chatbots to language education is their ability to provide immediate feedback. In conventional classrooms, it is often impossible for teachers to provide every learner with individual and continuous feedback. Chatbots, by responding immediately and individually to learner input, can support the learning process without interruption. In the context of TBLT, they can evaluate learners' linguistic performance while tasks are being carried out and provide guiding feedback where necessary. This feature increases learners' awareness of errors and develops their self-regulation skills.

Another important function of pedagogical agents in language education is to enhance motivation. Factors such as fear of making mistakes, performance anxiety, and social pressure can adversely affect learners' use of a foreign language. The non-judgemental nature of chatbots may enable learners to communicate more comfortably. Chatbot use can consequently contribute to more positive attitudes towards language learning, greater self-confidence, and reduced learning anxiety.

Despite the benefits of agents and chatbot technologies in language education, their use also presents several limitations. The responses generated by these systems may not always be accurate or pedagogically appropriate. Moreover, limited data relating to different languages and cultures may cause performance problems, particularly in languages other than English. Teacher oversight and pedagogical guidance are therefore of critical importance in AI-supported learning environments.

3. Gamification and Chatbots in the Language Learning Process

3.1. Gamification Elements for Enhancing Motivation and Reducing Language Anxiety

Success in foreign and second language education is determined not only by cognitive processes. Learners' affective characteristics also play a decisive role in learning outcomes. Motivation and foreign language anxiety therefore stand out as two fundamental affective variables that directly influence language learning. Anxiety experienced when using a foreign or second language may reduce learners' willingness to use the target language, restrict classroom participation, and adversely affect communicative performance. High levels of motivation, by contrast, enable learners to participate more actively, develop positive attitudes towards learning, and cope more effectively with learning-related problems. With recent advances in educational technology, gamification has attracted attention as an innovative instructional approach capable of both enhancing motivation and reducing foreign or second language anxiety. Gamification is defined as the transfer of game-design elements - including points, badges, levels, tasks, feedback mechanisms, competition, and collaboration - to non-game learning environments (Deterding et al., 2011, p. 10). These elements are considered capable of increasing learners' interest, making learning more enjoyable, and reducing negative emotions associated with foreign language use.

Foreign or second language anxiety refers to the tension, apprehension, and fear of failure experienced by learners when using the target language. Anxiety arising particularly during speaking activities may lead learners to

avoid communication and assume a passive role in the language learning process. Recent research indicates a negative relationship between foreign language anxiety and academic achievement. In a study of university learners of English, Özer and İşpınar Akçayoğlu (2021) found the following. Higher perceptions of foreign language self-efficacy were associated with lower anxiety and greater academic achievement.

Motivation and anxiety are dynamic constructs that frequently operate in an inverse relationship. A sense of success during learning increases learners' motivation while reducing anxiety. Whereas intense anxiety may weaken their willingness to learn and their self-confidence. Contemporary approaches to language teaching therefore focus not only on developing language skills but also on addressing learners' affective needs. This is one of the principal reasons for the growing prominence of gamification. By providing a sense of achievement and progress together with continuous feedback, gamified learning environments can make the learning process more motivating.

The effects of gamification on motivation are largely explained through Self-Determination Theory. According to this theory, the needs for competence, autonomy, and relatedness must be satisfied if motivation is to be sustained. In gamified learning environments, point systems, badges, and progression through levels support learners' sense of competence. Whereas the freedom to select tasks strengthens autonomy (Ryan & Deci, 2020). Teamwork and collaborative activities additionally enhance learners' sense of social belonging. Gamification thus offers a multidimensional structure that can support both intrinsic and extrinsic motivation.

A further factor in the capacity of gamification to reduce foreign language anxiety is its potential to lessen the fear of making mistakes. In conventional classrooms, learners may be reluctant to speak and experience performance anxiety because they are concerned about making errors. In gamified activities, however, errors are treated as a natural part of learning. Learners can repeat tasks, experiment with different strategies, and encounter no social sanctions when they are unsuccessful. This offers a significant advantage, particularly in reducing foreign language speaking anxiety.

Technological developments have contributed to the emergence of new learning environments that can further amplify the effects of gamification.

Virtual reality (VR), augmented reality (AR), and AI-supported learning environments, in particular, allow learners to participate in more immersive and realistic learning experiences. An experimental study by Kaplan-Rakowski and Gruber (2023) found the following. Speaking activities conducted in highly immersive virtual-reality environments significantly reduced learners' foreign language speaking anxiety. The researchers report that learners were less apprehensive about making mistakes in the virtual environment and were able to rehearse real-life situations in a safe setting.

Feedback mechanisms also play an important role in effective gamification. Receiving immediate feedback after performing tasks helps learners regulate their learning and develop a sense of achievement. Whereas feedback is frequently delayed in traditional instructional environments, gamified systems allow learners to see results immediately, making learning more transparent and motivating. Progress bars, leaderboards, and achievement indicators used in digital gamification applications enable learners to monitor their own development. Such practices can strengthen learners' perceptions of self-efficacy and foster positive feelings towards the learning process.

The social dimension of gamification likewise affects motivation and anxiety. Collaborative tasks, team-based competitions, and shared goals increase social interaction and strengthen learners' sense of belonging. Excessively competitive practices may, however, increase anxiety among some learners. A balanced use of competitive and collaborative elements is therefore recommended in gamification design. Systematic reviews indicate that collaborative gamification practices, in particular, can increase the motivation of foreign and second language learners while reducing their anxiety levels.

3.2. Providing Immediate and Personalised Feedback through Chatbots

Feedback is recognised as an essential component of learning. By making visible the discrepancy between a learner's current and target performance, it functions as a fundamental pedagogical tool that contributes to the regulation of learning (Özalp & Kaymakcı, 2022). Timely and personalised feedback, in particular, helps learners manage their learning more

effectively and develop self-regulation skills. Their capacity to identify strengths and weaknesses, recognise errors, and regulate learning depends largely on the nature of the feedback they receive. In conventional educational settings, however, time and workload substantially constrain teachers' ability to provide every learner with continuous, detailed, and personalised feedback. The increasing use of AI-supported chatbot technologies in education offers new responses to this problem. Chatbots are now employed not only as question-and-answer systems but also as intelligent learning tools that analyse learner performance, identify learning needs, and provide feedback suited to individual characteristics. One of the most important functions of educational chatbots is to provide immediate, personalised feedback on learning. These systems can identify areas of difficulty, track individual progress, and adapt learning processes to learner needs.

One of the principal advantages offered by chatbots concerns the timing of feedback. Whereas learners in conventional instructional settings often receive feedback on their errors only after a delay, chatbots can intervene immediately when errors arise during learning. Research demonstrates that feedback mechanisms support self-regulation and that self-regulation skills are positively related to academic achievement (Bilgili & Keklik, 2022). Confirmatory feedback provided in online learning environments has also been reported to relate significantly to self-regulation skills. Chatbots can provide learners with immediate guidance when they use an incorrect linguistic form, give an incomplete response, or fail to achieve the intended outcome during a task-based activity. This is particularly important in TBLT because learners can recognise linguistic gaps arising during meaningful communicative tasks without delay.

Personalised feedback is among the most prominent educational contributions of chatbot technologies. Although providing similar feedback to all learners is common in conventional classrooms, AI-supported systems can analyse learners' previous performance, learning histories, and patterns of error to provide differentiated feedback. This contribution is especially important in foreign and second language education, where learners have different linguistic strengths and weaknesses. By accounting for such individual differences, AI-supported chatbots can generate suggestions aligned

with a learner's needs. Performance-based recommendations, individual study plans, and adaptive learning content are among the principal features of these systems.

In language education, the feedback function of chatbots is not limited to error correction. Systems based on large language models, including ChatGPT, can provide detailed feedback across multiple dimensions, such as grammatical accuracy, lexical choice, text organisation, discourse structure, and communicative appropriateness. Such systems have been reported to be particularly effective in the development of writing skills. Karataş et al. (2024) found the following. ChatGPT contributed positively to foreign and second language learners' development in writing, grammar, and vocabulary while also increasing learning motivation and classroom engagement.

Within TBLT, chatbots can provide adaptive feedback on difficulties encountered during task performance. In a role-play activity, for example, they can evaluate whether a learner's expressions are appropriate to the intended communicative purpose, suggest ways of addressing omissions, and propose alternative formulations. This capability supports one of the central principles of the task-based approach. Learning through meaningful communication.

Alongside its benefits, chatbot-generated feedback has several limitations. AI systems may sometimes offer inaccurate, incomplete, or pedagogically inappropriate suggestions. Learners' over-reliance on these systems may also adversely affect the development of critical-thinking skills. Teacher oversight is therefore essential if chatbots are to be used effectively in instruction. Chatbots should be regarded not as replacements for teachers but as supplementary systems that support teaching and learning. Establishing human-AI collaboration in feedback processes will both enhance content accuracy and enable learners' pedagogical needs to be addressed more effectively.

4. Classroom Implementation Models

In TBLT, AI-supported chatbots can extend beyond individual out-of-class use. They can become natural and functional components of classroom activities. Contemporary technology-enhanced approaches recommend integrating digital tools directly into classroom ecosystems. Such tools should become integral to teaching and learning. In technology-mediated TBLT,

digital tools are not peripheral aids. They contribute actively to task design, performance, and assessment. Chatbots can provide continuous language input, interaction, and immediate feedback. They therefore create opportunities in both face-to-face and online environments.

The task-based approach acquires pedagogical value when translated into practicable classroom models. Chatbots can serve as a bridge between theoretical principles and classroom practice. Tasks in which learners attempt to solve real-life problems become more dynamic and learner-centred through the interactive affordances of chatbots. Chatbot support can diversify task-oriented communication, meaning-making, problem-solving, and decision-making, thereby enabling the core principles of TBLT to be implemented more effectively through technology. The use of chatbots in task-based practice allows tasks to be performed in more functional, authentic, and interactive ways.

An effective language learning environment requires a balance between learners' individual learning needs and their need for social interaction. Learners require opportunities to work at their own pace, repeat content, make mistakes without fear, and recognise and develop both weaker and stronger areas. At the same time, the inherently communicative and social nature of language necessitates environments that support interaction in social contexts. As flexible tools capable of addressing both requirements, chatbots can support individual study as well as group tasks, collaborative problem-solving, and peer interaction. They can therefore assist teachers with classroom activation and workload while supporting learners' individual learning and task performance.

This section examines the integration of chatbots into classroom instruction across three principal dimensions. It first considers individual activities and language practice that learners can undertake at their own pace, before discussing the functions of chatbots in group work and social learning. Finally, it considers how chatbots can promote not only learner-system interaction but also communication and peer learning among learners. The discussion thus illustrates the broad range of applications for chatbot-supported TBLT, extending from individual to collaborative learning.

4.1. Individual Learning Activities with Chatbot Support

Although conventional approaches often plan instruction on the assumption that all learners progress at the same rate, language development in fact varies in accordance with individual differences. Some learners acquire new forms and vocabulary rapidly, whereas others require more repetition and practice. Chatbots help accommodate these differences by providing interactive learning environments that can be adapted to individual needs. Through interaction with a chatbot, learners can determine their own pace, ask again about points they have not understood, and practise as much as necessary. This strengthens the learner-centred nature of TBLT and gives learning a more personalised character. Chatbots can make it possible to structure the process around the target group's age, first language, sociocultural characteristics, learning potential, pace, and strategy use, thereby increasing instructional effectiveness.

Foreign and second language learners may encounter a range of individual barriers. One important affective barrier is communication anxiety arising from the fear of making mistakes. In performance-based activities such as speaking and writing, learners may fear negative evaluation by the teacher or their peers and consequently avoid using the target language. Owing to their non-judgemental nature, chatbots can provide a digitally safe space in which learners may learn through errors without fear or anxiety. When learners produce an inaccurate expression, they can try again, formulate alternatives, and use the target language more freely without experiencing social pressure. Personalised activities may thus help particularly reticent learners and those with low self-confidence overcome their fear of errors and engage with the target language more comfortably.

A significant advantage of individual chatbot applications is their capacity to move grammar and vocabulary instruction beyond mechanical practice. Gamified micro-tasks created with chatbots provide opportunities to use target-language forms in meaningful contexts. Grammar instruction comprises three dimensions - form, meaning, and use - yet traditional language teaching frequently concentrates on form and may neglect functional use. A chatbot may, for example, ask a learner to explain a situation without using a particular word, solve a problem using specified linguistic forms, or complete short point-based tasks. Narrative task sequences and graded

levels of difficulty maintain attention while making repetition more natural and motivating. Grammar and vocabulary learning is thereby supported through tasks integrated with communicative purposes. Chatbot-supported activities allow target-language forms and lexis to be used naturally and internalised in terms of both meaning and function. Internalised language is more likely to be retained and used without communication breakdowns in accordance with sociolinguistic appropriateness.

The widespread use of chatbots in individual learning also transforms the teacher's role. Rather than functioning primarily as a transmitter of knowledge, the teacher becomes a facilitator who manages learning. While learners work individually with chatbots, the teacher can observe the classroom, identify those requiring support, and plan differentiated instruction. Performance data derived from chatbot interactions may also provide valuable information about learners' strengths and areas for development.

4.2. Chatbot-Supported Group Work, Social Learning, and Classroom Dynamics

Language learning should be understood not solely as an individual cognitive process, but as a process of meaning-making through social interaction. According to social constructionism, learning occurs through individuals' interactions with their environment, and language development rests to a considerable extent on communicative experience. Collaboration towards shared goals, the exchange of ideas, and the negotiation of meaning therefore occupy an important place in TBLT. Although AI-supported chatbots may initially appear to direct learners towards individual linguistic experiences, when appropriately designed they can perform a pedagogical function that supports group work and strengthens collaborative learning.

Chatbot functions in group work vary according to instructional purpose. A chatbot may guide learners by providing clues and additional information. It may instead moderate communication and manage game rules. In gamified scenarios, it can function as a non-player character (NPC). During travel planning, it may enact a hotel employee or tourist-information officer. It may also serve as a customer-service representative. In negotiation tasks, it may represent a customer or manager awaiting a solution. These applications increase task authenticity and motivational value. They

also expose learners to experiences resembling real-life communication. Moreover, they increase peer interaction and exposure to natural target-language use. They can strengthen peer learning and make instruction more active and enjoyable.

The success of group work in task-based learning depends not only on task quality but also on the clear definition of group members' responsibilities. Assigning different roles in chatbot-supported activities can make collaborative learning more systematic and help group members understand and adopt their responsibilities. One learner may, for example, communicate with the chatbot and formulate prompts, another may record the information provided, and a third may evaluate language use and alignment with task objectives. Such role allocation ensures active participation by all learners and increases individual accountability during task performance. Bringing together different perspectives can also foster more creative solutions to problems arising during the task. Group work can thereby become more than an activity in which only a small number of groups compete or discuss.

One important outcome of chatbot-supported group activities is their positive effect on classroom climate. Working towards common goals strengthens a sense of belonging among group members. While gamification elements can make learning more enjoyable and motivating. Intergroup task competitions, point systems, and collaborative problem-solving activities may increase classroom participation. Competition should nevertheless be prevented from becoming excessive and should be balanced with collaboration. Appropriately designed chatbot-supported tasks allow learners to experience both individual achievement and the social satisfaction of contributing to group success. This can help establish sustained motivation and encourage more active use of the target language.

4.3. Enhancing Peer Interaction through Chatbots

A common criticism of chatbot-supported language learning is the risk that learning will be confined to individual interaction between learner and system. One of the principal aims of TBLT. However, is to enable learners to communicate with other people through the target language. Chatbots should therefore be designed not only to support learner-system interaction

but also as mediating tools that create new opportunities for communication among learners. When chatbot-generated content, feedback, or tasks trigger discussion, decision-making, and negotiation among learners, the learning experience becomes more social and communicative. Technology thus ceases to separate learners and instead serves to enhance peer interaction. To achieve this aim, the teacher must structure instruction effectively in the role of facilitator.

Information-gap tasks constitute one effective means of enhancing peer interaction. In these tasks, each learner possesses only part of the information required to complete the task. A chatbot-supported application can provide different learners with different clues or pieces of data. In a scenario involving a tourist who is lost in a city, for example, the chatbot may give Learner A transport information and Learner B location information. To complete the task, the learners must share their information in the target language and reach a joint solution. In doing so, they actively use communication strategies such as asking questions, explaining, seeking confirmation, and resolving misunderstandings. The chatbot is transformed from a simple information provider into a mechanism that promotes the negotiation of meaning between learners. Its use is consequently not restricted to individual learning, as critics suggest, but can become an effective means of peer learning.

Chatbots can also support peer assessment. Peers may review productive performance undertaken with a chatbot. For example, a learner may present a chatbot-assisted story to classmates. Other learners can then evaluate its language, coherence, and alignment with task objectives. Learners may also discuss the accuracy or adequacy of chatbot feedback. These practices require both language production and analysis. They consequently develop metacognitive awareness. Learners can evaluate their own and others' learning more consciously. They can then regulate learning, develop strategies, and engage in more effective peer learning.

A particularly important contribution of chatbot-supported peer interaction is its capacity to develop the higher-level communication strategies required in real-life communication while enabling learners to complete tasks integrated with TBLT. During task performance, learners do not merely attempt to produce accurate linguistic forms. They also seek to understand

their interlocutors, resolve misunderstandings, request clarification, and construct shared meaning. This process, described in language-acquisition research as the negotiation of meaning, is critical to the development of communicative competence. By confronting learners with situations resembling natural communication, chatbot-supported peer tasks foster these abilities. They can also contribute to social skills such as respect for different views, empathy, and collaborative problem-solving. Chatbots thereby function not only as technological tools for language teaching but also as mediators of communicative and social learning.

5. Sample Chatbot-Supported Task-Based Activities and Game Designs

The pedagogical value of chatbots becomes more visible when they are translated into activities that can be implemented in the classroom. This section therefore presents sample activities and game designs in which chatbots are used in accordance with TBLT. The activities are structured around tasks that require learners to use the target language as a genuine means of communication.

The examples comprise scenarios targeting different language skills and assigning the chatbot a range of roles. In each application, the chatbot may function as a game master, a guide, or a character with whom learners interact. The aim is to move beyond using the chatbot solely as a question-and-answer tool and transform it into a learning partner that creates meaningful environments for communication.

5.1. Application I: Collaborative Story-Construction Task

Level: B1
Duration: 45-50 minutes
Mode of work: Groups of three or four
Core skills: Writing, speaking, and spoken interaction
Supporting skills: Vocabulary, sequencing events, problem-solving, and collaboration

1. Communicative Objective

Learners collaboratively construct a story in the target language. They establish cause-and-effect relationships between events. They explain why they agree or disagree with group suggestions. They also make joint decisions about the story's development. The task requires predictions, suggestions, expressions of possibility, and justification. Learners also recount events in the past.

2. Pre-task Preparation

The teacher divides learners into groups of three or four. Each group prepares a coherent written story and a brief oral presentation.

To prepare learners for the task, the teacher:

- reviews the elements of character, setting, time, event, problem, and resolution in a story;
- reviews sequencing expressions such as 'first', 'then', 'at that moment', 'suddenly', and 'finally';
- provides brief examples of the Turkish past tense, the inferential past tense, and basic structures used in storytelling;
- asks learners to identify vocabulary relating to characters, places, objects, and events that might be used in the story;
- explains that not all chatbot-generated information will necessarily be accurate or coherent and that the group must evaluate its outputs.

Group members may be assigned roles such as story writer, chatbot operator, coherence monitor, and presenter. In groups of three, one learner may assume more than one role.

3. Task Cycle

a. Initiating the story

The chatbot generates the beginning of a story appropriate to B1 level. For example:

> "Bir sabah okula gitmek için evden çıkan Ali, yol kenarında üzerinde yalnızca kendi adının yazılı olduğu eski bir kutu buldu. Kutunun üzerinde 'Kutuyu açmadan önce dikkatlice düşün.' yazıyordu."

(One morning, as Ali left home to go to school, he found an old box by the roadside with only his own name written on it. The box bore the message: Think carefully before opening.)

b. Developing the story collaboratively

Each group member takes a turn adding a new event to the story. The new event must connect with the preceding section and advance the narrative. As learners jointly decide what to add, they make suggestions, predictions, and justify their preferences in the target language.

At the end of each round, the chatbot:

- briefly summarises the story developed thus far;
- asks the group a question if there is an evident inconsistency between events;
- suggests a new problem, character, or unexpected event that could advance the story;
- offers options that require learners to decide rather than providing a solution directly.

Learners are not required to accept the chatbot's suggestion verbatim. They may accept it, modify it, or reject it while explaining their reasons.

c. Problem-solving and decision-making

The chatbot introduces a problem that must be solved. The key to the box might be lost, part of a map might be missing, or a section of the message in a letter might be unclear. Group members discuss possible solutions and decide how the story should continue.

d. Completing the story

Groups write a meaningful ending to their stories. The completed text should contain an opening, development, and conclusion; events should progress in a logical sequence; and the problem should be resolved.

e. Presentation and interaction

Each group presents its story orally to the class. After the presentation, other learners ask at least one question about the plot or characters. The presenting group answers on the basis of information contained in the story.

4. Tangible Task Outcome

By the end of the task, each group produces:

- a collaborative story comprising an opening, development, and conclusion;
- a brief plot outline showing the sequence of events;
- a three- to five-minute oral group presentation;
- a brief group reflection explaining which chatbot suggestions were accepted or rejected and why.

Using criteria announced in advance, class members may evaluate the stories in categories such as 'most coherent plot', 'most creative solution', or 'most effective use of the target language'.

5. Post-task Language Focus

Following the presentations, the teacher focuses on shared linguistic needs evident in the groups' texts. Rather than addressing every error that did not disrupt communication, this stage considers salient examples that affected task completion or text comprehensibility.

The language focus may examine:

- uses of the Turkish past tense and inferential past tense;
- sequencing expressions;
- consistent reference to people and objects;
- structures expressing cause-and-effect relationships;
- lexical repetition and contextually appropriate alternatives;
- direct and reported speech;
- spelling and punctuation.

Learners revise their stories in response to teacher and peer feedback and produce a corrected final version. Corrections proposed by the chatbot are also evaluated comparatively under teacher guidance.

6. Teacher and Chatbot Roles

Teacher roles:

- explaining the task objective, stages, and expected outcome;
- organising groups and allocating within-group roles;
- providing necessary linguistic support;

- encouraging interaction in the target language;
- checking that chatbot outputs are appropriate to the level and context and reflect natural Turkish usage;
- observing task performance and collecting assessment evidence;
- providing post-task linguistic feedback.

Chatbot roles:

- generating the opening situation;
- supporting the process as game master and narrator;
- summarising the story at the end of each round;
- offering options for new characters, problems, or unexpected events;
- asking thought-provoking questions rather than giving learners direct answers;
- suggesting words or expressions when requested by the group;
- offering suggestions concerning story coherence.

The chatbot is neither the final decision-maker nor the sole source of assessment. Learners and the teacher make the final decisions regarding story content and linguistic revisions.

7. Assessment Criteria

The task may be assessed against the following criteria:

- Task completion: Production of the required story and oral presentation.
- Plot coherence: Logical and interconnected progression of events.
- Communicative effectiveness: Clear expression of ideas and participation in within-group communication.
- Linguistic appropriateness: Use of vocabulary, grammar, and linking expressions appropriate to B1 level.
- Fluency and oral delivery: Clear, organised presentation of the story at an appropriate pace.
- Lexical range: Use of new and contextually appropriate vocabulary.
- Collaboration: Balanced participation by group members in decision-making and text construction.

- Creativity: Original development of the problem, events, and resolution.
- Critical use of AI: Questioning, modifying, or rejecting chatbot suggestions where necessary.

Assessment may combine teacher assessment, peer assessment, and group self-assessment.

8. Potential Chatbot Errors and Control Measures

During the activity, the chatbot may use vocabulary above B1 level, introduce inconsistencies between events, forget previous details, or generate expressions that do not reflect natural Turkish usage. It may also weaken the learner-centred nature of the task by writing a substantial portion of the story for the learners. To mitigate these risks:

- the prompt should clearly specify the proficiency level, task objective, and chatbot role;
- the chatbot should be instructed to generate only a brief summary and one problem in each round;
- an explicit constraint should state that the chatbot must not complete the story on the learners' behalf;
- the group should check every new output for consistency with preceding events;
- B1-level alternatives should be requested for vocabulary above the target level;
- unnatural or questionable expressions should be referred to the teacher;
- learners should not enter real names, addresses, school details, or similar personal information into the chatbot;
- chatbot output should not be accepted as the final text but reviewed and revised by learners.

Sample Prompt

> "Türkçeyi yabancı dil olarak öğrenen B1 düzeyindeki öğrenicilerle ortaklaşa öykü oluşturma görevi yürüteceksin. İlgi çekici ve B1 düzeyine uygun, 40–50 sözcüklük bir öykü başlangıcı oluştur. Öğrenicilerin yerine öyküyü tamamlama. Her turda öğrenicilerin yazdığı bölümü en fazla iki cümleyle özetle ve öyküyü ilerletecek yalnızca bir problem, sürpriz

olay veya düşündürücü soru sun. Önceki olayları ve karakter özelliklerini değiştirme. Dil bilgisi düzeltmelerini görev sırasında doğrudan yapma; yalnızca anlamayı engelleyen bir kullanım olduğunda açıklayıcı bir soru sor. Öğreniciler 'Görev tamamlandı.' yazdığında olay örgüsündeki olası tutarsızlıkları ve B1 düzeyinin üzerindeki ifadeleri ayrı başlıklar altında belirt."

(You will facilitate a collaborative story-construction task with B1 learners of Turkish as a foreign language. Generate an engaging 40-50-word story opening appropriate to B1 level. Do not complete the story on the learners' behalf. In each round, summarise the section written by the learners in no more than two sentences and offer only one problem, unexpected event, or thought-provoking question that will advance the story. Do not alter previous events or character traits. Do not directly correct grammar during the task; ask a clarifying question only when a usage impedes comprehension. When the learners write Task completed, identify possible plot inconsistencies and expressions above B1 level under separate headings.)

5.2. Application II: Contextual Vocabulary Inference and Reasoning Task

Level: A2

Duration: 35-40 minutes

Mode of work: Groups of three or four

Core skills: Reading comprehension, speaking, and spoken interaction

Supporting skills: Vocabulary, prediction, explanation, classification, and establishing cause-and-effect relationships

1. Communicative Objective

This task requires learners to infer word meanings from contextual clues, justify their inferences, and use the identified words for a specific real-life purpose.

Throughout the task, learners use communicative functions including:

- making predictions;
- expressing opinions;
- expressing agreement and disagreement;
- giving reasons;
- describing characteristics and functions;
- making comparisons;

- making joint decisions.

Each group identifies objects correctly and selects five for a new student flat. The group then prepares a justified 'Essential Items List'. Vocabulary inference thus becomes part of a real-life decision-making task.

2. Pre-task Preparation

The teacher presents a scenario in which the class must identify the basic needs of a person moving into a new student flat:

> "Arkadaşınız Türkiye'de yeni bir öğrenci evine taşındı. Evde çok az eşya var. Ona yardımcı olmak için evde gerekli olan nesneleri bulacak ve en önemli beş nesneyi seçeceksiniz."

(Your friend has moved into a new student flat in Türkiye. There are very few items in the flat. To help, you will identify necessary household objects and select the five most important ones.)

Before the task, the teacher:

- briefly reviews basic vocabulary relating to the home, kitchen, bathroom, and study area;
- models structures used to describe objects' shape, location, and function;
- reviews expressions such as 'I think it might be ...', 'Perhaps ...', 'Because ...', 'I agree/disagree with this', and 'This is more important';
- explains that naming the word alone is insufficient and that every prediction must be supported by at least one reason;
- explains that chatbot clues may suggest more than one object and that additional clues can be requested where necessary.

Group roles may include chatbot operator, prediction recorder, justification monitor, and group spokesperson.

3. Task Cycle

a. Selecting the hidden word

The chatbot selects a concrete A2-level object from the categories of kitchen, study area, bathroom, or everyday life. Without stating the word, it provides clues that progress from easier to more difficult.

Each clue may describe:

- where the object is found;
- its shape;
- its purpose;
- other objects used with it.

b. Group-based inference

After each clue, group members discuss possible answers and explain their predictions using structures such as:

- 'I think the answer might be ... because ...'
- 'I do not think it is ... because ...'
- 'I agree with my classmate because ...'
- 'We should ask for another clue. This clue is not sufficient.'
- 'This object is used for ...'
- 'This object is usually found in ...'

In each round, the group agrees on a shared prediction and submits it to the chatbot. This requires members to evaluate one another's views rather than respond individually or at random.

c. Questioning the clues

If a clue applies to more than one object, learners request a more discriminating clue from the chatbot. For example:

> "Bu ipucu hem bıçak hem de kaşık için uygun olabilir. Nesnenin biçimi veya nasıl kullanıldığı hakkında başka bir ipucu verir misin?"
>
> (This clue could apply to both a knife and a spoon. Could you provide another clue about the object's shape or how it is used?)

This stage requires learners to question chatbot output and request the information they need in the target language.

d. Using the word in context

Once the correct word has been identified, group members produce at least two sentences containing it. One sentence explains the object's function and the other describes a situation in which it is used in everyday life.

For example:

- "Kaşık yemek yerken kullanılan bir mutfak eşyasıdır." (A spoon is a kitchen utensil used when eating.)
- "Çorba içerken kaşık kullanırım." (I use a spoon when I eat soup.)

e. *Joint decision-making*

After identifying five or six words, groups select the five objects most essential for the new student flat. They rank them in order of importance and prepare a brief justification for each choice.

For example:

> "Birinci sıraya tencereyi koyduk; çünkü evde yemek hazırlamak için tencereye ihtiyaç vardır."
>
> (We ranked the saucepan first because a saucepan is needed to prepare food at home.)

f. *Presentation and comparison*

Each group presents its essential-items list to the class. Groups compare their lists and create a shared class list of five objects. An object must be justified by at least one group before it can be included.

4. Tangible Task Outcome

At the end of the task, each group prepares:

- a list of the words inferred;
- sample sentences explaining each word's characteristics and function;
- a five-item 'Essential Items List for a New Student Flat', arranged in order of importance;
- a brief oral presentation explaining why the selected objects are necessary;
- a shared class list produced by comparing the group lists.

5. Post-task Language Focus

After task completion, the teacher addresses shared linguistic needs evident in learners' predictions and justifications.

The language focus may examine:

- making predictions with Turkish expressions equivalent to 'I think', 'perhaps', and 'might';
- giving reasons with expressions equivalent to 'because' and 'therefore';
- expressing purpose with the Turkish structure için;
- expressing location with the Turkish locative construction -de/-da bulunur;
- explaining function with the construction -mek/-mak için kullanılır;
- adjectives used to describe objects;
- singular and plural forms;
- the accusative and locative case suffixes;
- comparative and superlative structures such as 'more important' and 'most essential'.

Learners revise sentences produced during the task in response to teacher and peer feedback. The class also evaluates chatbot-generated examples for meaning, naturalness, and appropriateness to A2 level.

6. Teacher and Chatbot Roles

Teacher roles:

- explaining the real-life scenario and tangible task outcome;
- providing limited pre-task support with necessary vocabulary and structures;
- observing participation and target-language use within groups;
- encouraging target-language use without simply prohibiting recourse to learners' first language;
- monitoring chatbot clues for accuracy, clarity, and level appropriateness;
- facilitating discussion where multiple answers are possible;
- providing post-task feedback on shared linguistic needs.

Chatbot roles:

- selecting concrete words appropriate to A2 level;
- generating clues that progress from easier to more difficult;
- asking learners to justify their predictions;
- providing further clarification for ambiguous clues;

- revealing the correct answer at an appropriate point;
- providing brief examples of the word in context;
- offering motivating but non-misleading feedback.

The chatbot should not evaluate a prediction merely as correct or incorrect. It should briefly explain why the prediction does or does not fit the clue provided.

7. Assessment Criteria

The task may be assessed against the following criteria:

- Task completion: Preparation of the essential-items list and oral presentation.
- Contextual inference: Accurate interpretation of clues.
- Justification of predictions: Clear and relevant reasons for predictions.
- Use of vocabulary in context: Production of meaningful and appropriate sentences with the identified words.
- Communicative effectiveness: Clear expression of opinions.
- Linguistic appropriateness: Use of vocabulary and structures appropriate to A2 level.
- Decision-making: Explanation of the ranking assigned to the selected objects.
- Collaboration: Participation by group members in discussion and decision-making.
- Critical chatbot use: Recognition of ambiguous or inaccurate clues and requests for clarification.

Teacher observation, group self-assessment, and peer assessment may be used in combination.

8. Potential Chatbot Errors and Control Measures

Chatbot clues may sometimes apply to more than one word. For example, the clue ‘It is found in the kitchen and is used when preparing food’ might suggest both a knife and a spoon. In such cases, a learner’s reasonable prediction should not simply be marked incorrect.

The chatbot may also:

- select words above A2 level;
- introduce abstract concepts, idioms, or culture-specific expressions prematurely;
- provide clues inconsistent with the word's meaning;
- repeat the same clue in different wording;
- reject a correct prediction;
- produce contradictions between clues;
- generate unnatural sample sentences.

To mitigate these risks:

- the prompt should specify that only common, concrete, everyday words appropriate to A2 level are to be selected;
- each clue should be required to remain consistent with earlier clues;
- plausible alternative predictions should be accepted;
- the chatbot should evaluate learners' reasoning before revealing the answer;
- teachers and learners should question ambiguous clues;
- idioms and abstract concepts should be excluded from this A2-level application;
- the teacher should check chatbot-generated sample sentences;
- learners should be reminded that chatbot evaluations are not necessarily definitive.

Sample Prompt

> "Türkçeyi yabancı dil olarak öğrenen A2 düzeyindeki öğreniciler için bağlamsal sözcük tahmin görevi yürüteceksin. Mutfak, çalışma odası, banyo veya günlük yaşam kategorilerinden somut ve yaygın bir nesne seç. Sözcüğü doğrudan söyleme. Kolaydan zora doğru ilerleyen, birbiriyle tutarlı ve A2 düzeyine uygun en fazla dört ipucu ver. Her ipucundan sonra öğrenicilerden ortak bir tahmin yapmalarını ve tahminlerini 'çünkü' kullanarak açıklamalarını iste. Bir tahmin verdiğin ipucuna uygun ancak hedeflediğin sözcükten farklıysa tahmini doğrudan yanlış sayma; neden mümkün olduğunu belirt ve daha ayırt edici bir ipucu ver. Doğru sözcük bulunduğunda nesnenin işlevini açıklayan A2 düzeyinde bir örnek cümle iste. Öğrenicilerin yerine cümle kurma ve cevabı erken açıklama."

(You will facilitate a contextual vocabulary-inference task for A2 learners of Turkish as a foreign language. Select a common, concrete object from the categories of kitchen, study area, bathroom, or everyday life. Do not state the word. Provide no more than four mutually consistent A2-level clues, progressing from easier to more difficult. After each clue, ask learners to agree on a prediction and justify it using because. If a prediction fits your clue but differs from your intended word, do not simply mark it incorrect; explain why it is plausible and provide a more discriminating clue. Once the correct word has been identified, ask for an A2-level example sentence explaining the object's function. Do not formulate the sentence for the learners or reveal the answer prematurely.)

Sample Interaction

Chatbot: Yeni bir öğrenci evi için gerekli olabilecek gizli bir nesne seçtim. İlk ipucu: Bu nesne genellikle mutfakta bulunur.

Öğrenici 1: Bence bu nesne tabak olabilir. Çünkü tabak mutfakta bulunur.

Chatbot: Tahminin ilk ipucuna uygun; ancak hedeflediğim nesne tabak değil. İkinci ipucu: Bu nesneyi genellikle elimizle tutarak kullanırız.

Öğrenici 2: Bıçak olabilir. Çünkü bıçağı elimizle tutarız ve mutfakta kullanırız.

Chatbot: Bu da mantıklı bir tahmin. Ancak ipuçları henüz bıçakla başka bazı mutfak eşyalarını ayırmak için yeterli değil. Üçüncü ipucu: Bu nesneyi yemek yerken kullanırız. Genellikle çorba ve yoğurt gibi yiyeceklerle birlikte kullanılır.

Öğrenici 3: Bence cevap kaşık. Çünkü kaşıkla çorba ve yoğurt yeriz.

Chatbot: Evet, doğru cevap "kaşık". Şimdi "kaşık" sözcüğünü kullanarak bu nesnenin işlevini açıklayan bir cümle kurun.

Öğrenici 4: Kaşık, çorba içmek ve yemek yemek için kullanılır.

Chatbot: Cümleniz anlaşılır ve sözcüğü doğru bağlamda kullandınız. Şimdi grubunuzla karar verin: Kaşık, yeni bir öğrenci evi için hazırlayacağınız temel ihtiyaç listesine girmeli mi? Neden?

(Chatbot: I have selected a hidden object that may be needed in a new student flat. First clue: This object is usually found in the kitchen.

Learner 1: I think it might be a plate because plates are found in the kitchen.

Chatbot: Your prediction fits the first clue, but my intended object is not a plate. Second clue: We usually hold this object in our hand when using it.

Learner 2: It might be a knife because we hold a knife in our hand and use it in the kitchen.

Chatbot: That is also a reasonable prediction. The clues are not yet sufficient to distinguish a knife from several other kitchen utensils. Third clue: We use this object when eating, usually with foods such as soup and yoghurt.

Learner 3: I think the answer is a spoon because we eat soup and yoghurt with a spoon.

Chatbot: Yes, the correct answer is 'spoon'. Now use the word 'spoon' in a sentence explaining the object's function.

Learner 4: A spoon is used for eating soup and other food.

Chatbot: Your sentence is clear and uses the word in the correct context. Now decide as a group: Should a spoon be included in the essential-items list for a new student flat? Why?)

5.3. Application III: Role-Play Task Based on Real-Life Problems

Level: B1
Duration: 45-50 minutes
Mode of work: Groups of two or three
Core skills: Speaking and spoken interaction
Supporting skills: Listening, problem-solving, requesting information, negotiating, and decision-making

To develop listening skills, the chatbot's voice-conversation function should be used or its written responses should be rendered through a text-to-speech tool. If interaction is exclusively text-based, the target skills should instead be specified as reading, writing, and written interaction.

1. Communicative Objective

This task requires learners to report a lost-luggage problem that they might encounter in real life, request necessary information, and negotiate with an employee in order to resolve the problem.

Throughout the task, learners use communicative functions including:

- explaining a problem;
- recounting a past event;
- describing objects and belongings;
- requesting and providing information;
- requesting clarification;
- expressing possibility;
- proposing solutions;
- making requests;
- deciding between alternatives;

and related communicative functions.

By the end of the task, learners are expected to complete a lost-luggage report, select a delivery option, and explain the solution agreed with the employee.

2. Pre-task Preparation

The teacher presents the following real-life scenario: 'You have arrived in Ankara from Berlin via Istanbul, but cannot find your luggage in the baggage-reclaim area. You must go to the lost-luggage office, explain the situation, provide the necessary information, and agree with the employee on how your luggage will be delivered to you.'

Before the task, the teacher:

- reviews basic vocabulary relating to airports, flights, and luggage;
- reviews adjectives describing colour, size, shape, material, and distinguishing features;
- briefly models tense structures used to recount past events;
- addresses formulaic expressions for requesting information and making requests;
- explains that real personal information must not be entered into the chatbot;
- distributes role cards containing entirely fictional flight, luggage-tag, telephone, and address details.

The following expressions may be reviewed before the task:

- "Bagajımı teslim alanında bulamadım."
- "Bagajımın nerede olduğunu kontrol edebilir misiniz?"

- "Bagajımı tarif etmem gerekiyor mu?"
- "Ne zaman teslim edilebilir?"
- "Adresime gönderilmesini istiyorum."
- "Başka bir seçeneğiniz var mı?"
- "Bu çözüm benim için uygun değil; çünkü …"
- "Bagaj bulunduğunda bana nasıl bilgi vereceksiniz?"
- 'I could not find my luggage in the baggage-reclaim area.'
- 'Could you check where my luggage is?'
- 'Do I need to describe my luggage?'
- 'When can it be delivered?'
- 'I would like it to be sent to my address.'
- 'Is there another option?'
- 'This solution is not suitable for me because ...'
- 'How will you inform me when the luggage is found?'

3. Task Cycle

a. Establishing roles and the information gap

Learners receive role cards containing different information.

Learner 1 - Luggage owner: Knows the flight information, basic characteristics of the luggage, and luggage-tag number.

Learner 2 - Travel companion: Knows the luggage's distinguishing features, some of its contents, and delivery requirements.

Chatbot - Lost-luggage employee: Requests the information required for the report, explains possible circumstances, and presents solution options.

Distributing information between group members requires learners to listen to one another and exchange information in order to complete the task.

b. Explaining the problem

Learners explain why they have contacted the lost-luggage office. The chatbot asks for necessary information such as the flight number, departure and arrival points, transfer details, and luggage-tag number.

Rather than asking all questions at once, the chatbot should ask them sequentially in a manner consistent with the natural flow of the conversation.

c. Describing the luggage

Learners describe the luggage's:

- type;
- size;
- colour;
- material;
- brand or label;
- distinguishing features.

If the information is insufficient, the chatbot asks clarifying questions. For example:

- "Bavulunuz sert mi yoksa kumaş mı?"
- "Üzerinde ayırt edici bir işaret var mı?"
- "Yaklaşık olarak ne kadar büyüktü?"
- 'Is your suitcase hard-shell or fabric?'
- 'Does it have a distinguishing mark?'
- 'Approximately how large was it?'

d. Gathering information and negotiating

The chatbot provides information about the luggage's possible status and offers different solutions, such as:

- collecting the luggage from the airport once it has been found;
- having the luggage delivered to the accommodation address;
- receiving information about procedures for immediate essential needs.

Learners discuss the advantages and limitations of the options. If an unsuitable solution is offered, they explain why and request an alternative.

For example:

> "Havaalanına tekrar gelmemiz zor; çünkü şehir merkezine uzak bir yerde kalıyoruz. Bagajın adresimize gönderilmesi mümkün mü?"
>
> (It would be difficult for us to return to the airport because we are staying far from the city centre. Could the luggage be delivered to our address?)

e. Reaching a joint solution

The learners and chatbot agree on:

- the delivery method;
- the anticipated notification process;
- the fictional contact details to be used;
- the tracking number.

The task ends when the chatbot provides a fictional lost-luggage reference number and the learners accurately summarise the agreed solution.

f. Reporting and classroom comparison

At the end of the task, each group briefly explains to the class:

- the nature of the problem;
- the information it provided;
- the options it considered;
- the solution it selected;
- the communication difficulties encountered during the interaction.

Groups compare the communication and negotiation strategies they employed.

4. Tangible Task Outcome

At the end of the task, each group produces:

- a lost-luggage report form completed with fictional information;
- a brief description of the luggage;
- the delivery and notification plan agreed during the interaction;
- a fictional lost-luggage tracking number;
- a three- to five-minute oral group presentation explaining how the problem was resolved;
- a brief self-assessment of the communication and negotiation strategies used.

Successful task completion depends not only on linguistic accuracy but also on learners' ability to convey the necessary information and reach a practicable solution.

5. Post-task Language Focus

After task completion, the teacher addresses shared linguistic needs that emerged during the interactions.

The language focus may examine:

- recounting events with the witnessed past tense in Turkish;
- expressing possibility with the Turkish construction "-mIş olabilir";
- adjectives denoting colour, shape, size, and material;
- the accusative, dative, locative, and ablative case suffixes;
- making polite requests and expressing possibility with the construction "-Ebilmek";
- the verbs "istemek, talep etmek, and tercih etmek";
- question words and indirect-question forms;
- giving reasons with expressions equivalent to "çünkü ('because), bu nedenle ('therefore), ancak (however), and bu yüzden (for this reason)";
- forms of address, requests, and expressions of thanks used in formal interaction.

Without interrupting task performance, the teacher records observations and subsequently shares selected examples with the class. Learners jointly correct problematic expressions and generate alternatives serving the same function.

Expressions used by the chatbot are also evaluated for naturalness, politeness, contextual appropriateness, and suitability to B1 level.

6. Teacher and Chatbot Roles

Teacher roles:

- explaining the real-life problem and task outcome;
- designing the role cards and information gap;
- providing limited pre-task support with necessary vocabulary and communicative formulae;
- checking audio quality and intelligibility when voice interaction is used;
- monitoring balanced participation during the task;

- checking chatbot-provided information for accuracy and realism;
- preventing learners from sharing real personal information;
- taking observation notes during the task;
- providing post-task linguistic and communicative feedback.

Chatbot roles:

- enacting the lost-luggage office employee;
- requesting necessary information in a natural conversational sequence;
- asking clarifying questions about incomplete or ambiguous information;
- providing level-appropriate explanations of the luggage's possible status;
- offering more than one solution;
- allowing learners opportunities to decide and negotiate;
- summarising the agreed solution at the end of the interaction;
- generating a fictional report or tracking number.

The chatbot should not solve the problem for learners, generate all information itself, or dominate the interaction with long, one-sided explanations.

7. Assessment Criteria

The task may be assessed against the following criteria:

- Task completion: Completion of the lost-luggage report with the necessary information and attainment of a practicable solution.
- Information transfer: Clear and comprehensible communication of flight and luggage information.
- Object description: Description of the luggage's distinguishing features in sufficient detail.
- Requesting information: Use of appropriate questions to obtain necessary information.
- Problem-solving: Evaluation of the options offered and selection of a suitable solution.
- Negotiation: Justification of requests and solicitation of alternatives where necessary.
- Communicative effectiveness: Clear and purposeful communication of the message.

- Spoken interaction: Turn-taking, listening to the interlocutor, and responding appropriately.
- Linguistic appropriateness: Use of vocabulary and structures appropriate to B1 level.
- Fluency: Ability to sustain speech without serious interruption to communication.
- Collaboration: Participation by group members in information sharing and decision-making.
- Digital safety: Use of fictional data rather than real personal information.

Teacher observation, the task outcome, group self-assessment, and peer feedback may be used in combination.

8. Potential Chatbot Errors and Control Measures

The chatbot may provide unrealistic information about airline procedures, refer to rights or services that do not exist, or state with certainty when the luggage will be found. It may also forget previously supplied information, repeat questions, or use technical language above B1 level. Other risks include:

- requesting real personal information;
- presenting the fictional scenario as a genuine registration process;
- completing form fields on the learners' behalf;
- confusing flight and luggage information;
- producing excessively long and unnatural employee responses;
- using forms of address inappropriate to formal communication;
- rejecting a reasonable learner solution without justification;
- assuming that listening practice has occurred when interaction is text-based.

To mitigate these risks:

- the prompt should state that all names, flights, addresses, and contact details are fictional;
- sharing real names, telephone numbers, addresses, passport details, or ticket information should be prohibited;

- it should be made explicit that the chatbot has no access to a real airline system;
- the chatbot should explain possible situations and options rather than assert definite outcomes;
- it should ask no more than one or two questions in each turn;
- it should be instructed to retain previously supplied information throughout the interaction;
- technical vocabulary above B1 level should be explained or replaced with simpler alternatives;
- the teacher should not present the chatbot's legal or institutional explanations as factual information;
- teachers and learners should evaluate the final solution for realism.

Sample Prompt

> "Türkçeyi yabancı dil olarak öğrenen B1 düzeyindeki öğrenicilerle kayıp bagaj problemine dayalı bir rol yapma görevi yürüteceksin. Sen havalimanındaki kayıp bagaj ofisi görevlisisin. Öğreniciler, Berlin'den İstanbul aktarmalı olarak Ankara'ya gelen ve bagajlarını teslim alanında bulamayan iki yolcuyu canlandıracak. Bütün kişi, uçuş, bagaj, telefon ve adres bilgileri kurgudur. Gerçek kişisel bilgi isteme. Görüşmenin başında problemi öğren ve uçuş bilgisi, bagaj etiket numarası, bagajın özellikleri ve teslimat tercihi hakkında soruları doğal bir sırayla yönelt. Her turda en fazla iki kısa soru sor. Öğrenicilerin yerine cevap verme ve sorunu hemen çözme. Eksik bilgi bulunduğunda açıklayıcı soru sor. Bagajın havaalanından alınması veya kurgu konaklama adresine gönderilmesi gibi en az iki çözüm seçeneği sun ve öğrenicilerden birini gerekçelendirerek seçmelerini iste. B1 düzeyine uygun, doğal ve nazik Türkçe kullan. Kesin teslim zamanı veya gerçek bir havayolu garantisi verme. Görüşmenin sonunda kararlaştırılan çözümü özetle ve açıkça kurgu olduğu anlaşılan bir takip numarası oluştur."

> (You will facilitate a role-play task based on a lost-luggage problem with B1 learners of Turkish as a foreign language. You are an employee at the airport lost-luggage office. The learners will enact two passengers who have arrived in Ankara from Berlin via Istanbul and cannot find their luggage in the baggage-reclaim area. All names, flight details, luggage details, telephone numbers, and addresses are fictional. Do not request real personal information. At the beginning, identify the problem, then ask questions about flight information, the luggage-tag number, luggage characteristics, and delivery preferences in a natural sequence. Ask no more than two short questions in each turn. Do not

answer for the learners or resolve the problem immediately. Ask clarifying questions when information is missing. Offer at least two solutions, such as collection from the airport or delivery to a fictional accommodation address, and ask learners to select and justify one. Use natural, polite Turkish appropriate to B1 level. Do not provide a definite delivery time or a real-airline guarantee. At the end, summarise the agreed solution and generate a tracking number that is clearly fictional.)

Sample Interaction

Scenario: Two learners have arrived in Ankara from Berlin via Istanbul but cannot find their luggage in the baggage-reclaim area. The chatbot enacts an employee at the airport lost-luggage office. All information used in the task is fictional.

Learner roles:

- Learner 1 - Ayşe: The passenger whose luggage is missing; knows the flight and luggage-tag information.
- Learner 2 - Mehmet: Ayşe's travel companion; knows the luggage's distinguishing features and delivery requirements.
- Chatbot: Lost-luggage office employee.

Chatbot (Görevli): Merhaba, kayıp bagaj ofisine hoş geldiniz. Size nasıl yardımcı olabilirim?

Ayşe: Merhaba. Berlin'den İstanbul aktarmalı olarak Ankara'ya geldik ancak bagajımı teslim alanında bulamadım.

Chatbot: Geçmiş olsun. Uçuş numaranızı ve bagaj etiketi numaranızı söyleyebilir misiniz? Lütfen yalnızca görev kartınızdaki kurgu bilgileri kullanın.

Ayşe: Uçuş numaramız TK1234, bagaj etiketi numaramız ise BG458732.

Chatbot: Teşekkür ederim. Bagajınızı tarif edebilir misiniz? Rengi, büyüklüğü ve ayırt edici bir özelliği var mı?

Ayşe: Büyük ve siyah bir bavul. Ön tarafında kırmızı bir etiket var.

Mehmet: Ayrıca bavulun tutacağına mavi bir kurdele bağladık. Bavul sert malzemeden yapılmış.

Chatbot: Verdiğiniz bilgiler yeterince açık. Kurgu kayıt sistemine göre bagajınızın İstanbul'daki aktarma sırasında uçağa yüklenmemiş olması mümkündür. Bagaj bulunduğunda Ankara Havalimanı'ndan alabilir veya konaklama adresinize gönderilmesini isteyebilirsiniz. Hangisini tercih edersiniz?

Ayşe: Havaalanına tekrar gelmek bizim için zor olabilir. Bagajın kaldığımız yere gönderilmesi mümkün mü?

Chatbot: Evet, görev kapsamında kurgu konaklama adresinize gönderilmesini talep edebilirsiniz. Ancak teslimat zamanı kesin değildir. Bu seçeneği neden tercih ettiğinizi kayda eklememi ister misiniz?

Mehmet: Evet. Şehir merkezinde kalıyoruz ve havaalanına tekrar gelmek zaman alacak. Bu nedenle adrese teslim edilmesini istiyoruz.

Chatbot: Anladım. Bagaj bulunduğunda kurgu telefon numaranıza bilgi verilecek ve bagaj belirttiğiniz kurgu adrese gönderilecek. Kararlaştırdığımız çözüm bu mudur?

Ayşe: Evet. Bagajımız bulunduğunda bize bilgi verilmesini ve kaldığımız yere gönderilmesini istiyoruz.

Chatbot: Bildiriminiz oluşturuldu. Kurgu takip numaranız "EĞİTİM-BG-2048"dir. Bu numara yalnızca sınıf içi görev için geçerlidir. Başka bir konuda bilgi almak ister misiniz?

Mehmet: Hayır, teşekkür ederiz. Bagaj bulunduğunda yaklaşık ne kadar süre içinde bilgi alabiliriz?

Chatbot: Kesin bir süre veremem; ancak bagajın durumu güncellendiğinde kurgu iletişim numaranıza bilgi gönderilecektir.

Ayşe: Tamam. Yardımınız için teşekkür ederiz.

Chatbot: Rica ederim. Umarım bagajınız kısa sürede bulunur. İyi günler dilerim.

(Chatbot (Employee): Hello, welcome to the lost-luggage office. How can I help you?

Ayşe: Hello. We arrived in Ankara from Berlin via Istanbul, but I could not find my luggage in the baggage-reclaim area.

Chatbot: I am sorry to hear that. Could you give me your flight number and luggage-tag number? Please use only the fictional information on your role card.

Ayşe: Our flight number is TK1234 and the luggage-tag number is BG458732.

Chatbot: Thank you. Could you describe your luggage? What colour and size is it, and does it have any distinguishing features?

Ayşe: It is a large black suitcase with a red label on the front.

Mehmet: We also tied a blue ribbon to the handle. The suitcase has a hard shell.

Chatbot: The information you have provided is sufficiently clear. According to the fictional system, your luggage may not have been loaded onto the aircraft during the transfer in Istanbul. Once it is found, you may collect it from Ankara Airport or request delivery to your accommodation address. Which would you prefer?

Ayşe: It might be difficult for us to return to the airport. Could it be sent to where we are staying?

Chatbot: Yes. For the purposes of this task, you may request delivery to your fictional accommodation address, although the delivery time cannot be guaranteed. Would you like me to record why you prefer this option?

Mehmet: Yes. We are staying in the city centre and returning to the airport would take time. We would therefore like it delivered to the address.

Chatbot: Understood. When the luggage is found, a notification will be sent to your fictional telephone number and the luggage will be delivered to the fictional address you provided. Is that the solution we have agreed?

Ayşe: Yes. We would like to be informed when our luggage is found and have it sent to where we are staying.

Chatbot: Your report has been created. Your fictional tracking number is 'TRAINING-BG-2048'. This number is valid only for the classroom task. Is there anything else you would like to know?

Mehmet: No, thank you. Approximately how long will it take to hear from you once the luggage is found?

Chatbot: I cannot provide an exact time, but a notification will be sent to the fictional contact number when the luggage status is updated.

Ayşe: All right. Thank you for your help.

Chatbot: You are welcome. I hope your luggage is found soon. Have a good day.)

6. Challenges in Classroom Implementation and Proposed Solutions

Chatbot-supported TBLT offers applications ranging from individual study to collaborative activities and gamified tasks. As with any technological innovation, however, introducing AI-supported tools into classrooms creates challenges. Their pedagogical value depends not only on their affordances but also on anticipating implementation barriers, risks, and limitations. Risks associated with large language models include misinformation, bias, and over-reliance (Kasneci et al., 2023; Martunyuk, 2024). A learning process that is poorly structured or inaccessible to some learners may cause more harm than benefit.

Challenges in chatbot-supported practice fall into three broad categories. The first comprises infrastructural barriers involving internet access, devices, and technical equipment (UNESCO, 2025). The second concerns output accuracy, reliability, and pedagogical effects (Labadze et al., 2023). The third involves data privacy, ethical responsibility, and developmentally appropriate tools (Berson et al., 2025). These dimensions should be considered holistically and relationally (Bardakcı, 2024). Inadequate infrastructure may deepen inequality. Erroneous output may cause inaccurate learning. Ethical violations may threaten learner safety and privacy. Chatbots should therefore supplement teachers rather than replace them (Huang et al., 2022).

This section examines technological and infrastructural barriers, pedagogical risks including inaccurate language, hallucinations, and loss of focus, and issues of digital ethics, data privacy, and age-appropriate chatbot selection. Possible solutions are presented under each heading.

6.1. Technological and Infrastructural Barriers: Internet Access and Device Shortages

Classroom implementation depends on uninterrupted internet access and sufficient digital devices. Where these conditions are absent, technology is not equally accessible and a digital divide emerges. Rofiah et al. (2022) identified unstable connections and inadequate device access as major factors adversely affecting participation and learning experience. Economic and technological inequalities may similarly restrict access to AI-based

tools and deepen the digital divide. The target group's sociocultural circumstances are therefore highly consequential.

Experience of remote education demonstrates the importance of infrastructure. Ülger (2021) found the following. Inadequate internet access, device shortages, and the sharing of a single device, particularly in larger families, severely restricted participation. Learners without a device or connection may likewise be excluded from individual or group-based chatbot tasks.

Infrastructure problems also include connection failures, device malfunctions, server delays, and limited school bandwidth, all of which can disrupt activities and waste lesson time. Teachers' digital competence is equally important. Even where equipment is available, teachers without sufficient experience of managing chatbots pedagogically may encounter difficulties.

Low-bandwidth designs that can also operate offline or asynchronously may reduce access problems; asynchronous strategies have been recommended in response to the limitations of synchronous practice (Rofiah et al., 2022). Station-based arrangements with shared devices can support group participation where devices are scarce. Strengthening digital infrastructure, providing device and connectivity support, and improving teachers' technical preparedness are further institutional measures (Ülger, 2021).

6.2. Pedagogical Risks: Inaccurate Linguistic Forms, Hallucinations, and Loss of Focus

Chatbots may generate inaccurate or pedagogically inappropriate language. Because large language models rely on probabilistic patterns rather than semantic understanding, they may produce grammatically flawed, contextually inappropriate, or level-inappropriate expressions. Kohnke et al. (2023) identify misinformation, bias, and cultural insensitivity among their limitations and emphasise the need for teacher oversight. Learners' internalisation of erroneous forms may adversely affect language development.

A major risk is hallucination. The generation of factually incorrect, unverifiable, or fabricated content in fluent and persuasive language. Zhai et al. (2024) note that misleading content complicates learners' evaluation of information. Its authoritative presentation may encourage uncritical acceptance, particularly among learners who lack sufficient target-language proficiency to evaluate output accuracy.

These risks may be more pronounced in languages other than English. The predominance of English in training data may increase morphological and syntactic errors, semantic shifts, and unnatural expressions in comparatively under-represented languages such as Turkish. Yağcı and Yıldız (2023) found ChatGPT outputs not consistently accurate or coherent, reinforcing the importance of critical evaluation, teacher oversight, and output verification.

Although AI systems can provide rapid and formally consistent feedback, they may struggle to explain an error's context, prioritise errors, or adapt feedback to individual needs. Chatbot feedback therefore cannot fully replace context-sensitive guidance grounded in teachers' pedagogical judgement. Superficial or inaccurate feedback may cause confusion and reduce motivation.

Open-ended chatbot interaction may also lead learners away from the task and its objective. Irrelevant topics or responses can disrupt activities. In gamified designs, points, badges, and competition may shift attention from language learning to the game itself, causing gamification to become an end rather than a means.

Another risk is over-reliance and cognitive offloading. Delegating thinking, production, and problem-solving to chatbots may impair critical thinking and self-regulation. Zhai et al. (2024) report possible weakening of decision-making, critical-thinking, and analytical skills. While Gerlich (2025) identifies a negative relationship between frequent AI use and critical thinking through cognitive offloading, particularly among younger users. In language education, this may inhibit learners' independent language production.

These risks can be managed through teacher supervision and human-AI collaboration. Teachers should review outputs, correct misleading content, and encourage critical evaluation. AI literacy and habits of questioning and verification are essential (Karataş et al., 2024). Well-designed prompts and tasks with explicit objectives can preserve instructional focus, and pedagogical purposes should remain central to activity planning (Guo et al., 2025).

6.3. Digital Ethics, Data Privacy, and Age-Appropriate Chatbot Selection

AI systems collect substantial data through user interaction. Learners' conversations, texts, and interaction data may be recorded. While collection, storage, and use often remain unclear. Temur (2025) identifies data privacy

and security as priority ethical concerns and notes the serious issues raised by collecting and storing learner data.

Insufficient transparency in the data policies of free educational chatbots creates risks concerning sensitive data. Data privacy, algorithmic bias, transparency, and equity require careful evaluation (Varol, 2025). Although Türkiye's Personal Data Protection Law (KVKK) and the EU General Data Protection Regulation (GDPR) establish principles for processing personal data, extensive cross-border data flows can complicate their application (Temur, 2025).

Ethical issues extend beyond privacy. Bias in training data may produce inadequate or stereotypical representations of cultural, social, or linguistic groups. Chatbot use in assignments also raises questions of originality and academic integrity. Use should therefore be structured around transparency, accountability, and fairness (Varol, 2025).

Age-appropriate selection is equally important. Children and adolescents may be less able to evaluate AI limitations and more inclined to perceive systems as trustworthy human-like entities. Kurian (2025) highlights an 'empathy gap'. Systems simulate empathy without genuine emotional understanding and may produce inappropriate responses. Because services impose different age and parental-consent requirements, classroom tools must suit learners' age and developmental level.

Educational settings should prioritise age-appropriate, privacy-assuring tools developed where possible for education. Data minimisation, informed consent from learners and where necessary parents or guardians, and explicit institutional AI policies can reduce ethical risks (Temur, 2025). Teachers and parents should supervise younger learners' interactions (Kurian, 2025). AI literacy and digital-ethics awareness can enable conscious, safe, and responsible use within a framework protecting learner rights and safety.

7. Conclusion and Recommendations

Rapid advances in artificial intelligence have encouraged educational restructuring and innovative language-teaching practices. Large-language-model chatbots offer interactive, accessible, and personalisable learning experiences. This chapter has examined their theoretical and practical

integration with TBLT. It has also considered gamification, motivation, language anxiety, and interaction. Classroom activities and games have illustrated the practicability of chatbot-supported instruction.

Chatbots support TBLT by enabling learners to solve real-life problems, reason about meaning, participate in role-play, and perform collaborative tasks. Immediate feedback, performance-responsive content, and adaptability across proficiency levels make them powerful pedagogical tools. In teaching Turkish as a foreign or second language, they offer opportunities for speaking practice, written production, vocabulary instruction, and intercultural interaction.

Gamification makes learning more engaging and supports intrinsic motivation through points, badges, levels, tasks, and narrative framing. Integrated with chatbots, it provides continuous feedback, a sense of achievement, and reduced fear of mistakes, potentially lowering foreign language anxiety, strengthening learner autonomy, and increasing active participation.

Nevertheless, chatbot-supported instruction has limitations, including inaccurate or pedagogically inappropriate content, cultural bias, and over-reliance. Chatbots should be treated as pedagogical supports rather than teacher replacements. Effective integration requires balanced human-AI collaboration.

Chatbot-supported task-based practice remains an emerging research area in teaching Turkish as a foreign or second language. Whereas much existing research concerns English teaching, experimental studies in Turkish remain limited, highlighting the need for further work.

The following recommendations may be proposed:

- More experimental studies should examine the effects of chatbot-supported tasks on language skills, motivation, self-regulation, and language anxiety.
- Chatbot systems specific to teaching Turkish as a foreign or second language should be developed, taking account of Turkish linguistic and cultural characteristics.
- Training should develop teachers' AI literacy and effective prompt-design skills to support pedagogical chatbot use.

- Applications should diversify integrated tasks and games for receptive and productive skills. Listening, speaking, reading, and writing.
- Gamification should balance competition with collaboration and enhance motivation without increasing anxiety.

In conclusion, integrating AI-supported chatbots with TBLT can create more interactive, learner-centred, and personalised environments for teaching Turkish as a foreign language. Their role is likely to expand as AI develops;. However, centring pedagogical principles, ethical responsibilities, and teacher guidance will maximise educational value. Chatbots can thus function not merely as technological innovations but as pedagogical tools supporting meaningful learning.

References

Bardakcı, S. (2024). Yapay zekâ-eğitim ilişkisine bütüncül bakış: Bir bilim haritalama çalışması. *Uluslararası Türk Eğitim Bilimleri Dergisi*, 12(3), 10801114.

Berson, I. R., Berson, M. J. & Luo, W. (2025). Innovating responsibly: Ethical considerations for AI in early childhood education. *AI, Brain and Child*, 1(1), 2.

Bilgili, H. ve Keklik, İ. (2022). Akademik Öz Yetkinlik ile Akademik Öz Düzenleme Arasındaki İlişkide Akademik Motivasyonun Aracı Rolü. *Yükseköğretim ve Bilim Dergisi,* 12(1), 213-227.

Çangal, Ö., Çelik, M. E. ve Başar, U. (2025). Yabancılara Türkçe öğretiminde yapay zekâ kullanımına yönelik öğretici görüşleri. *Aydın Tömer Dil Dergisi*, 10(1), 57–97.

Deterding, S., Dixon, D., Khaled, R. & Nacke, L. (2011). From game design elements to gamefulness: Defining gamification. In A. Lugmayr, H. Franssila, C. Safran, & I. Hammouda (Eds.), Proceedings of the 15th International *Academic MindTrek Conference: Envisioning Future Media Environments* (pp. 9–15). ACM.

Ellis, R. (2003). *Task-based language learning and teaching*. Oxford University Press.

Genç Ersoy, B. ve Ersoy, M. (2021). Teknoloji destekli Türkçe öğretimi üzerine yayınlanan makalelerin içerik analizi: ULAKBİM-TR Dizin

örneği. *Ahi Evran Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 7(3), 810–829.

Gerlich, M. (2025). AI tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies*, 15(1), Article 6.

Guo, K., Zhong, Y., Qin, J., & Chu, S. K. W. (2025). Investigating EFL teachers' lesson planning for chatbot-assisted learning of argumentative writing: a TPACK perspective. *Interactive Learning Environments*, 33(2), 1520-1541.

Huang, W. J., Hew, K. F. & Fryer, L. K. (2022). Chatbots for language learning—Are they really useful? A systematic review of chatbot-supported language learning. *Journal of Computer Assisted Learning*, 38(1), 237–257.

Kaplan-Rakowski, R. & Gruber, A. (2023). The impact of high-immersion virtual reality on foreign language anxiety. *Smart Learning Environments*, 10(46).

Karataş, F., Abedi, F. Y., Ozek Gunyel F., Karadeniz, D. ve Kuzgun, Y. (2024). An investigation into ChatGPT's effect on foreign language learners. *Education and Information Technologies*.

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., Stadler, M., Weller, J., & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274.

Kohnke, L., Moorhouse, B. L. & Zou, D. (2023). ChatGPT for Language Teaching and Learning. *RELC Journal*, 54(2), 537-550.

Kurian, N. (2025). 'No, Alexa, no!': Designing child-safe AI and protecting children from the risks of the 'empathy gap' in large language models. *Learning, Media and Technology,* 50(4), 621–634.

Labadze, L., Grigolia, M. & Machaidze, L. (2023). Role of AI chatbots in education: Systematic literature review. *International Journal of Educational Technology in Higher Education*, 20, Article 56.

Long, M. H. (2015). *Second language acquisition and task-based language teaching*. Wiley Blackwell.

Lume, L. L. & Hisbullah, M. (2022). The effectiveness of task-based language teaching to teach speaking skills. *Journal of Languages and Language Teaching*, 10(1), 85–93.

Martunyuk, O. (2024). Aı tools ın foreıgn language teachıng: pros and cons of ukraınıan and foreıgn practıces. *Comparative Professional Pedagogy*, 14(2), 65-73.

Michel, M. (2023). Key concepts in applied linguistics: Task-based language teaching. *ELIA*, 23, 317-328.

Mulyadi, D., Singh, C. K. S., Setiawan, A., & Prasetyanti, D. C. (2023). Technology-enhanced task-based language teaching toward self-directed language learning: ESP learners' views. *Studies in English Language and Education*, 10(3), 1326–1341.

Nget, S., Pansri, O. & Poohongthong, C. (2020). The Effect of Task-Based Instruction in Improving the English Speaking Skills of Ninth-Graders. *LEARN Journal: Language Education and Acquisition Research Network Journal*, 3(2), 208–224.

Noroozi, M. & Taheri, S. (2022). Task-based language assessment: A compatible approach to assess the efficacy of task-based language teaching vs. present, practice, produce. *Cogent Education*, 9(1), 2105775.

Nunan, D. (2004). *Task-based language teaching*. Cambridge University Press.

Özalp, M. T. ve Kaymakcı, S. (2022). Türkiye'de Eğitim Bilimleri Alanında Geri Bildirim Üzerine Yapılmış Lisansüstü Tezlerin Analizi. *Ahmet Keleşoğlu Eğitim Fakültesi Dergisi*, 4(2), 468-483.

Özer, Ö. ve İşpınar Akçayoğlu, D. (2021). Examining the roles of self-efficacy beliefs, self-regulated learning and foreign language anxiety in the academic achievement of tertiary EFL learners. *Participatory Educational Research*, 8(2), 357–372.

Rofiah, N. L., Aba Sha'ar, M. Y. M., & Waluyo, B. (2022). Digital divide and factors affecting English synchronous learning during COVID-19 in Thailand. *International Journal of Instruction,* 15(1), 633–652.

Ryan, R. M. & Deci, E. L. (2020). Intrinsic and extrinsic motivation from a self-determination theory perspective: Definitions, theory, practices, and future directions. *Contemporary Educational Psychology*, 61, 101860.

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford University Press.

Şengül, K. (2021). Uzaktan eğitim sürecinde Türkçenin yabancı dil olarak öğretimi. *RumeliDE Dil ve Edebiyat Araştırmaları Dergisi*, 24, 174–222.

Temur, S. (2025). Eğitimde yapay zekâ kullanımı: Etik sorunlar ve çözümler. *Mehmet Akif Ersoy Üniversitesi Eğitim Fakültesi Dergisi*, 74, 568–595.

UNESCO. (2025). AI and education: Protecting the rights of learners. *UNESCO*. https://unesdoc.unesco.org/ark:/48223/pf0000395373

Ülger, K. (2021). Uzaktan eğitim modelinde karşılaşılan sorunlar-fırsatlar ve çözüm önerileri. *Uluslararası Güncel Eğitim Araştırmaları Dergisi*, 7(1), 393–412.

Varol, N. (2025). Eğitimde yapay zekâ ve etik: Bir alanyazın taraması. *Okul Yönetimi*, 5(1), 117–125.

Wang, N. & Lynch, M. P (2022). Creating effective goals in task-based language teaching for online collaborative EFL writing tasks. *Pädagogische Horizonte*, 6(2), 85–105.

Willis, D., & Willis, J. (2007). *Doing task-based teaching*. Oxford University Press.

Yağcı, Ş. Ç. ve Yıldız, T. A. (2023). ChatGPT, yabancı dil öğrenicisinin güvenilir yapay zekâ sohbet arkadaşı mıdır? RumeliDE Dil ve Edebiyat Araştırmaları Dergisi, (37), 1315-1333.

Zhai, C., Wibowo, S. & Li, L. D. (2024). The effects of over-reliance on AI dialogue systems on students' cognitive abilities: A systematic review. *Smart Learning Environments*, 11, Article 28.

A Legal Analysis of the Use of Artificial Intelligence in Türkiye: Risk, Liability and Regulatory Approach

Alkın Karakurumer[1]

1. Introduction

Artificial intelligence technologies present themselves today not merely as a field of technical innovation, but as a structural and multi-layered element of transformation that transforms the norm-setting capacity of law, the instruments of intervention available to public authority and the relationship that individuals establish with fundamental rights. The spread of algorithmic decision-making processes strains the limits of the classical understanding of liability and makes it necessary to reinterpret fundamental legal concepts such as fault, causal link and foreseeability in the context of artificial intelligence.

The Bill on Artificial Intelligence, submitted to the Grand National Assembly of Türkiye in order to respond to this transformation, contains provisions on the definition of artificial intelligence systems, the determination of user and developer responsibilities, the prohibition of discriminatory data sets, the strengthening of the principles of transparency and human oversight, and the fight against deepfake content. The bill carries particular importance in that it is the first comprehensive and thematic piece of regulation prepared with the aim of filling the gap in the field of artificial intelligence.

This study addresses the bill in question in the light of the established principles of Turkish law, the practice of the high courts, academic studies and international regulatory trends, and examines comprehensively the legal

[1] Lecturer Dr., National Defence University, ORCID: 0000-0002-4415-0719, *alkin.karakurumer@msu.edu.tr*, Türkiye

and strategic consequences that the regulation may produce in terms of its normative consistency, its applicability and the balance of rights and freedoms.

The level of sophistication that artificial intelligence technologies have reached in recent years has turned these systems from auxiliary tools used in particular sectors into a structural infrastructure element that permeates the whole of social life. Algorithms fed by large data sets and continuously developing through machine learning and deep learning techniques are radically transforming not only technical processes but also the mechanisms of decision-making, classification, risk prediction and behavioural steering (Floridi vd., 2018). This transformation directly affects the norm-setting capacity of law and its classical conceptual framework.

In particular, the spread of artificial intelligence-supported systems in the exercise of public power has brought onto the agenda a new administrative practice referred to as "algorithmic administration". The use of algorithms in areas such as the allocation of social assistance, risk analyses, the selection of public personnel, and inspection and surveillance activities gives rise to serious debates concerning the limits of the administration's discretionary power, the obligation to give reasoned decisions and the scope of judicial review. In this context, artificial intelligence has become a horizontal element of transformation that directly affects not only private law relationships but also the fields of administrative law, constitutional law and criminal law.

This transformation also strains the classical assumptions of law that place human will at the centre. In algorithmic decision-making processes, the question of at which stage, by whom and to what extent an unlawful outcome can be foreseen is becoming increasingly complex. The self-learning capacity of artificial intelligence systems blurs the classical "instrument–perpetrator" distinction and brings onto the agenda the need, in the determination of legal liability, for an assessment that is not only outcome-oriented but process-oriented, encompassing the processes of design, training, data selection and oversight.

For this reason, contemporary states treat artificial intelligence not merely as a part of economic competitiveness or digital transformation policies, but as a strategic field of regulation in terms of public order, fundamental

rights and freedoms, democracy and the principle of the rule of law (Demir, 2020; Akıncı, 2021). Regulations concerning artificial intelligence also contain normative choices that redefine the position of the state vis-à-vis the individual.

In Türkiye, the legal framework concerning artificial intelligence applications was for a long time shaped by provisions of legislation of a general nature, in particular within the framework of the protection of personal data, freedom of expression and administrative law. However, this fragmented structure has not allowed the risks specific to artificial intelligence to be addressed in a systematic and holistic manner. (Gözler, 2020).

In the Turkish legal literature, a significant part of the academic debate on artificial intelligence has developed through postgraduate theses as well as peer-reviewed journal articles. These academic studies offer detailed and in-depth analyses of the legal nature of artificial intelligence, the relationship between algorithmic decision-making processes and human will, the limits of liability in terms of criminal and administrative law, and the effects on fundamental rights (Acar, 2018; Yılmaz, 2020; Koç, 2021; Arslan, 2022). In this respect, the thesis literature plays a central role in the formation of the conceptual infrastructure of artificial intelligence law in Türkiye.

2. Purpose, Method and Contribution of the Study

This study does not confine itself to treating the Bill on Artificial Intelligence merely as a normative text; it aims to evaluate the regulation in a multidimensional manner in the context of the principle of the rule of law, judicial review, comparative law and its potential for implementation. In the study, the method of doctrinal analysis has been adopted; the national and international literature, academic studies, and the case law of the Constitutional Court and the Court of Cassation have been examined systematically.

The principal contribution of the study is that it carries the risks arising from artificial intelligence beyond abstract ethical debates and analyses them through concrete legal problems such as user and developer liability, algorithmic administrative acts, content removal and access-blocking mechanisms, discriminatory data sets, and transparency and explainability obligations. In addition, how the case law of the Turkish high courts may

be applied by analogy to artificial intelligence applications is set out within a critical and holistic framework.

In this respect, the study does not offer merely a descriptive analysis explaining the existing regulation; it also aims to make visible the legal risks and implementation problems that the regulation may give rise to.

3. The Bill on Artificial Intelligence, the Article Justifications and the General Justification

3.1. The Bill

With regard to artificial intelligence, the principal articles, regulatory areas and possible legal effects of the Bill on Artificial Intelligence dated 07 November 2025, signed by Halil ÖZTÜRK, Attorney at Law, Member of Parliament for Kırıkkale of the Nationalist Movement Party (MHP), are summarised in Table 1. This table shows that the normative orientation of the bill is concentrated along the axes of risk, liability and oversight.

Table 1: The Principal Provisions of the Bill on Artificial Intelligence and Their Legal Effects

Article	Regulatory Area	Principal Provision Introduced	Normative Purpose	Legal Effect and Area of Debate
Art. 1	Definition	Artificial intelligence systems are defined in a technology-neutral and broad manner.	Preventing technologies from falling outside the scope	Risk of vagueness in terms of the principles of legal certainty and foreseeability
Art. 2	Criminal liability	Where artificial intelligence is directed to commit an offence, the user is the perpetrator; an increased penalty is envisaged for the developer.	Tying liability to human will	Technical difficulty in proving fault and the causal link
Art. 3	Content removal	Access blocking within 6 hours and joint liability for content originating from artificial intelligence	Rapid intervention and prevention of victimisation	Risk of weakening of proportionality and judicial review

Article	Regulatory Area	Principal Provision Introduced	Normative Purpose	Legal Effect and Area of Debate
Art. 4	Data sets	The use of discriminatory data sets is deemed an explicit violation.	Protection of equality and data security	Problem of technical detection and auditing of discrimination
Art. 5	BTK powers	Emergency access-blocking power for election security and public order	Preventive protection of public order	Expansion of the administration's discretionary power
Art. 6	Cybersecurity	Transparency, hallucination prevention and human oversight obligations	Safe and auditable artificial intelligence	Uncertainty as to the limits of technical obligations
Art. 7	Deepfake	Labelling of content generated by artificial intelligence is made mandatory.	Combating disinformation	Chilling effect on freedom of expression
Art. 8	Access blocking	Offences committed by means of artificial intelligence are brought within the scope of access blocking.	Alignment with the existing system	Risk of automatic and generalised intervention
Art. 9	Platform liability	The provisions of the Turkish Penal Code (TCK) are applied to artificial intelligence-based social networks.	Extension of platform liability	Cross-border enforcement and jurisdiction problems
Arts. 10–11	Entry into force	Provisions on entry into force and execution	Normative operability	Need for secondary legislation

When the provisions summarised in Table 1 are evaluated together, it is seen that the Bill on Artificial Intelligence carries the aim of addressing the risks arising from artificial intelligence within a holistic framework rather than through fragmented and technical interventions. Rather than positioning artificial intelligence systems as an independent legal subject, the bill places at its centre the areas of human-originated responsibility relating to the design, training and use of these systems. While this approach is generally consistent with the principles of fault, personal responsibility and imputability in Turkish law, it also brings with it problems of certainty, proportionality and judicial review that may arise in practice. Indeed, the areas

of legal debate indicated in the table make visible not only the normative choices of the regulation but also the fundamental points of tension concerning how these choices will take concrete shape in the processes of administrative implementation and judicial review. For this reason, the bill's risk-based regulatory approach, its liability regime and the limits of interference with fundamental rights will be examined in detail in Sections 3 and 4 of the study.

3.2. Article Justifications

"ARTICLE JUSTIFICATIONS

ARTICLE 1- The concept of artificial intelligence is defined with the aim of clarifying its legal framework. In this way, uncertainties in the application of legislation are prevented and the criteria under which artificial intelligence-based systems will be evaluated are clearly set out.

ARTICLE 2- With this provision added to the Turkish Penal Code, the liability regime for certain offences committed by means of artificial intelligence systems is regulated. In this way, it is aimed to prevent victimisation by clarifying the responsibilities of users and developers.

ARTICLE 3- In order to ensure the rapid removal of content that is contrary to personality rights, threatens public order or is of a deepfake nature, the access-blocking period is shortened and liability is extended. In this way, the prevention of victimisation is targeted.

ARTICLE 4- It is aimed to bring the data sets used in artificial intelligence applications into conformity with the prohibition of discrimination and with data security. By deeming the use of discriminatory data sets an explicit violation, the fundamental rights of individuals are protected.

ARTICLE 5- In order to prevent threats to election security and public order, the Information and Communication Technologies Authority (BTK) is granted the power of emergency intervention. Within this scope, the effectiveness of the regulation is increased by providing for deterrent administrative fines.

ARTICLE 6- In order to increase security in artificial intelligence systems and to prevent manipulative content, additional obligations are imposed on service providers. In addition, administrative sanctions are regulated for violations.

ARTICLE 7- It is aimed to combat disinformation by making it mandatory that fake content such as deepfakes be clearly indicated to the user. Public security is protected by providing for administrative sanctions for violations.

ARTICLE 8- New offences that may be committed with artificial intelligence systems are added to the offences within the scope of access blocking in the internet environment, and the access-blocking mechanism is strengthened.

ARTICLE 9- The area of liability is extended by ensuring that the provisions of the Turkish Penal Code are also applicable to artificial intelligence-based social network providers.

ARTICLE 10- It is the entry-into-force article.

ARTICLE 11- It is the execution article." (TBMM, 2025).

3.3. General Justification

"GENERAL JUSTIFICATION

Humanity, which is trying to adapt to and even to manage the great change and transformation in communication and information technologies taking place today at a dizzying pace, is encountering new, multidimensional and unknown issues at a level unseen in history. This wind of change has very serious and even destructive effects at the level of individuals, societies and even states, and forces all actors to act proactively, to produce rapid and innovative solutions, and to take bold steps and develop strategies. Being able to be proactive is a matter of being prepared, making long-term plans and addressing issues from the perspective of strategic management.

Artificial intelligence is described as a powerful technology that provides a machine with the ability to perform cognitive functions such as perceiving, reasoning and learning. Artificial intelligence makes it possible for machines to learn from experience, to adapt to new inputs and to perform tasks similar to those performed by humans. Scenarios such as the ethical dilemmas that artificial intelligence will bring with it, the transformation of social life into digital individuals, problems relating to rights and intellectual property, and the creation of artificial intelligence-based systems by intelligent systems are also being discussed by various researchers.

At this point, the first "National Artificial Intelligence Strategy", prepared by the Digital Transformation Office of the Presidency and the Ministry of Industry and Technology, which will guide Türkiye's work on artificial intelligence, entered into force upon publication in the Official Gazette dated 20/08/2021 and numbered 31574, and was shared with the public on 24 August 2021 at Bilişim Vadisi. The National Artificial Intelligence Strategy has been drawn up so as to cover the years 2021-2025.

> It should not be forgotten that if artificial intelligence, as a field, is not handled correctly, the risks may far outweigh its benefits. Technology entrepreneurs such as Facebook CEO Mark Zuckerberg argue that artificial intelligence can save human lives but can also be misused, and that it is therefore necessary to be very careful and cautious. On the other hand, Tesla and SpaceX CEO Elon Musk has suggested that artificial intelligence is probably one of the greatest existential threats facing humanity and argues that regulators need to be proactive.
>
> With the development of artificial intelligence technologies, the risks emerging in the digital environment have diversified, and violations of personality rights, threats to public order, insult, threat and the production of unethical content have become widespread. In particular, hate speech, deepfake content and content produced for the purpose of manipulation by means of artificial intelligence have become a serious security risk. In our country, it has become imperative to establish the necessary legal framework by amending the existing basic laws, without the need to enact a new law. It is our most fundamental duty to protect the rights and liberties, honour and dignity of the person and individual fundamental rights and freedoms, and to put into effect regulations to this end. On this basis, this regulation is of importance against the adverse situations and effects that artificial intelligence creates and will create." (TBMM, 2025).

4. The General Approach and Regulatory Logic of the Bill on Artificial Intelligence

The general approach of the Bill on Artificial Intelligence goes beyond a purely technical norm-setting effort. The regulation offers a normative and value-based framework that aims to balance the social effects of artificial intelligence, and seeks to strike a delicate balance between the promotion of technological innovation and the protection of fundamental rights. In this respect, the bill adopts neither a permissive approach based on technological determinism nor a rigid and prohibitive interventionism.

The backbone of the bill is formed by the principles of transparency, accountability and human oversight. These principles are not merely ethical objectives; they are also the concrete reflections of the principle of the rule of law in the digital age. A similar normative approach is seen to have been adopted in the European Union's Proposal for an Artificial Intelligence Regulation (AI Act) (European Commission, 2021; Yıldız, 2022).

4.1. The Risk-Based Regulatory Approach

The bill proceeds from the assumption that not every artificial intelligence application carries the same level of risk. For this reason, instead of absolute prohibitions, different obligations are envisaged according to the nature, scope and social impact area of the risk. This approach shows a large degree of parallelism with the European Union's risk-based regulatory model.

It is emphasised in particular that algorithmic systems used by public power should be subjected to stricter administrative and judicial review because of their potential to interfere directly with the fundamental rights of individuals (Bayraktar, 2021; Yurtseven, 2021). This approach reflects a dynamic interpretation of the principle of the rule of law that adapts to technological developments, rather than a static one.

The risk-based regulatory approach also makes regulatory flexibility possible. However, this flexibility must not be achieved at the expense of legal certainty and foreseeability. Otherwise, the vagueness of the concept of risk may lead to an excessive expansion of the administration's discretionary power.

The risk-based regulatory approach rests on the assumption that artificial intelligence applications may create social impact at different levels, and differentiates regulatory intervention according to this level of impact. At the same time, the integration of algorithmic systems into administrative decision-making processes leads to the emergence of decision mechanisms that operate at the system level rather than through the individual discretion of the administration. While this transformation carries the potential to increase administrative effectiveness, it also brings with it the risk of an indirect weakening of the mechanisms of constitutional review and accountability. For this reason, the risk-based regulatory approach must rest not only on technical classifications but also on normative guarantees aimed at preserving administrative responsibility and judicial review (Bovens & Zouridis, 2002).

4.2. A Comparative Law Perspective

Viewed from the perspective of comparative law, there are marked differences between the European Union's preventive and comprehensive regulatory model, the sectoral, fragmented and ex post intervention-based approach of the USA, and China's strict control system centred on public

order and political oversight. The bill in Türkiye is seen to be positioned closer to the European Union model in terms of its normative structure and the concepts used.

This choice is of a nature that may produce strategic consequences not only in terms of legal harmonisation but also in terms of international data flows, the digital economy and Türkiye's position in the global artificial intelligence ecosystem. However, rather than a one-to-one transfer of the European Union model, an adaptation compatible with Türkiye's constitutional structure and judicial practice is imperative.

The conceptual structure and risk-based approach of the bill in Türkiye show marked parallels with the European Union Artificial Intelligence Regulation. At the same time, it is pointed out in the comparative literature that this model harbours certain structural problems in terms of certainty and foreseeability. In particular, the classification of high-risk artificial intelligence systems and the legal consequences produced by this classification create an area that may expand the administration's discretionary power. For this reason, the assessments to the effect that, rather than a direct transfer of the European Union model, an adaptation compatible with the constitutional principles and judicial practice of domestic law is imperative should be taken into account (Veale & Borgesius, 2021).

4.3. A Comparative Analysis of the United Nations General Assembly's Resolution on Artificial Intelligence and Türkiye's Artificial Intelligence Bill

The resolution on artificial intelligence adopted by the United Nations (UN) General Assembly on 21 March 2024 stands out as one of the first comprehensive multilateral texts emphasising that, at the international level, artificial intelligence applications should be guided within the framework of the principles of human rights, security, transparency and auditability. The resolution calls on UN member states and all stakeholders to develop safe, secure and trustworthy artificial intelligence systems and, at the same time, to increase the contribution of artificial intelligence to the 2030 Sustainable Development Goals (Birleşmiş Milletler, 2024).

By stating that the rights afforded to people in online environments must be protected throughout the entire life cycle of artificial intelligence

systems, the UN Resolution proposes a human rights-based governance framework for national regulations (Birleşmiş Milletler, 2024).

The table below shows the fundamental principles, the forms of implementation, and the points of alignment and difference between the UN Resolution on Artificial Intelligence and Türkiye's Artificial Intelligence Bill.

Table 2: A Comparative Analysis of the UN Resolution on Artificial Intelligence and Türkiye's Artificial Intelligence Bill

Subject / Principle	**UN Resolution on Artificial Intelligence (21.03.2024)**	**Türkiye's Artificial Intelligence Bill (07.11.2025)**	**Alignment / Difference**
Transparency and human oversight	It is recommended that artificial intelligence systems be transparent and auditable, and that there be human supervision.	It is envisaged that human oversight be mandatory in public administration and critical systems, and that algorithms be audited (Art. 6).	Aligned; the Bill transforms the UN principles into binding provisions in domestic law.
Protection of fundamental rights and prohibition of discrimination	The protection of the fundamental rights of individuals and the prevention of discrimination are a priority.	The use of discriminatory data sets is prohibited; data security and equality are ensured (Art. 4).	Highly aligned; the Bill gives concrete form to the human rights-focused approach of the UN resolution.
Risk-based regulation	It is stated that preventive measures are to be taken according to the risk levels of artificial intelligence systems.	Strict administrative and judicial review is envisaged for high-risk systems (Section 3.1).	Aligned; international recommendations have been taken into account at the national scale.
Deepfake and disinformation	Attention is drawn to the risk of false information and manipulation in digital environments.	The labelling of deepfake content and a rapid access-blocking mechanism have been regulated (Art. 7, Section 3.5).	Aligned; the objective is the same, and the method has been given concrete form through national implementation.

Subject / Principle	UN Resolution on Artificial Intelligence (21.03.2024)	Türkiye's Artificial Intelligence Bill (07.11.2025)	Alignment / Difference
International cooperation & strategic approach	International standards and cooperation are emphasised.	A normative approach aligned with the European Union Artificial Intelligence Regulation is displayed; it is adapted to domestic law and constitutional principles (Section 3.2).	Aligned; international and national frameworks have been balanced.

4.4. The Definition of the Artificial Intelligence System and Legal Certainty

The definition of the artificial intelligence system introduced in the bill offers a framework broad and flexible enough to adapt to technological developments. Formulating the definition in this way aims to prevent new technologies from falling outside the regulation in the short term. At the same time, it should not be overlooked that this breadth harbours significant risks in terms of the principle of legal certainty.

The principle of legal certainty requires that individuals be able to foresee in advance, to a reasonable degree, which acts will produce legal consequences. It has frequently been argued that, especially where administrative sanctions, access-blocking decisions and the liability regime are concerned, vague definitions expand the administration's discretionary power and may undermine the principle of proportionality (Yurtseven, 2021; Bayraktar, 2021).

The importance of the principle of certainty is also clearly recognised in judicial practice. The Constitutional Court has held in many of its decisions that vague and ambiguous concepts in regulations concerning the digital domain may create a chilling effect on freedom of expression and legal security. (Anayasa Mahkemesi, 2015). In this context, giving concrete form to the definition of artificial intelligence through secondary legislation and technical guidelines appears imperative from the standpoint of normative consistency.

4.5. User and Developer Liability

One of the most critical aspects of the bill is that it places the legal liability of artificial intelligence systems not on an independent subject but on the axis of user and developer. The bill does not accept artificial intelligence as an autonomous perpetrator entirely detached from human will; it positions it as an instrument designed, trained and used by humans. This approach follows a line consistent with the principles of fault, personal responsibility and imputability in Turkish criminal law (Kara, 2019; Özgenç, 2020).

In academic studies too, there is a strong consensus that recognising artificial intelligence as an independent subject of criminal liability is incompatible with the existing systematics of criminal law (Koç, 2021; Arslan, 2022). These studies argue that liability should be directed at the user and the developer within the framework of criteria such as design defects in the software, failure to eliminate foreseeable risks and breach of the duty of supervision. However, it is also clear that this approach may give rise to significant difficulties in practice. Particularly in complex machine learning models, it may not always be technically possible to determine at which stage the harmful outcome emerged. This situation strains the limits of the classical understanding of liability in terms of proving fault and establishing the causal link.

The bill's avoidance of positioning artificial intelligence as an independent legal subject is apposite from the standpoint of the existing systematics of criminal law. Evaluating harm arising from artificial intelligence within the framework of the liability of the human actors who design, train and use the system is important for preserving the principles of fault and personal responsibility. At the same time, the views that classical fault-based liability may prove insufficient, especially in high-risk systems with limited foreseeability, and that objectivised liability models or models based on a reversed burden of proof should therefore be opened up for discussion are gaining increasing weight in the comparative criminal law literature (Gless, Silverman & Weigend, 2016).

4.6. Deepfake Content and the Limits of Intervention

The inclusion in the bill of specific provisions directed at deepfake content reveals the aspect of the regulation that takes current risks into account. It is clear

that the artificial intelligence-supported production of fake images and audio poses serious threats, in particular to personality rights, the privacy of private life and election security.

In Turkish judicial practice, the principles of proportionality, necessity and proportionality in the narrow sense are the fundamental reference points regarding intervention in digital content. In the decision of the Court of Cassation, it is emphasised that interference with freedom of expression can be considered legitimate only where a clear and serious violation of a right is at issue (Yargıtay 4. HD, 2017).

Academic studies likewise state that content created through automated generation cannot be entirely removed from the protection of freedom of expression, but that intervention is inevitable where deepfake technology is misused (Öztürk, 2020; Sevinç, 2023). In this context, it is of great importance that the rapid intervention mechanisms introduced by the bill be operated in a manner open to judicial review and subject to the requirement of a reasoned decision.

4.7. Discriminatory Data Sets and Data Security

The provisions in the bill concerning the prohibition of discriminatory data sets strengthen the fundamental rights-oriented aspect of the artificial intelligence regulation. It is now widely accepted that the data sets on which artificial intelligence systems are trained have the potential to reproduce social biases.

The case law of the Constitutional Court on the principle of equality and the protection of personal data constitutes the constitutional basis of this provision (Anayasa Mahkemesi, 2015). It is further stated that artificial intelligence not only reflects discrimination but can also systematically deepen it through algorithmic processes (Demirtaş, 2019; Şahin, 2022).

For this reason, the detection and auditing of discriminatory data sets and, where necessary, the withdrawal of the system from use should be regarded not merely as a technical matter but as a constitutional necessity.

4.8. The Principle of Transparency and Explainability

Although the principles of transparency and explainability are not regulated under an explicit article heading in the bill, they present themselves as fundamental principles permeating the regulation as a whole. It is strongly voiced that, particularly in algorithmic systems used by public administration, the

"black box" character is incompatible with the principle of the rule of law (Yılmaz, 2020; Bayraktar, 2021).

It is imperative for the effectiveness of judicial review that the algorithms used by the administration be explainable, at least with respect to the fundamental criteria on which the decision rests. Otherwise, the right of individuals to an effective remedy against administrative acts may become meaningless in practice.

4.9. Algorithmic Administration and the Principle of the Rule of Law

The integration of artificial intelligence-supported systems into administrative decision-making processes brings with it a qualitative transformation in the classical understanding of administrative law. This transformation is expressed in the literature through the concepts of "algorithmic administration" or "algorithmic governance", and means the transfer of the discretionary power that the administration exercises through individual public officials to largely automated decision mechanisms operating at the system level (Bovens & Zouridis, 2002; Bayraktar, 2021; Yurtseven, 2021). While algorithmic administration has the potential to increase administrative effectiveness and the speed of decision-making, it also raises serious normative problems with regard to the fundamental components of the principle of the rule of law.

The principle of the rule of law requires that all acts and actions of the administration be bound by law, that arbitrariness be prevented, and that individuals have effective remedies against administrative acts. As also emphasised in the case law of the Constitutional Court, the limitation of the administration's powers by the principles of certainty, foreseeability and proportionality is the fundamental condition of legal security (Anayasa Mahkemesi, 2015). With respect to administrative acts established through algorithmic systems, however, preserving this principle gives rise to a need for review that is different from, and more complex than, that for classical administrative acts.

This is because algorithmic decision-making processes often have the character of a "black box" far removed from transparency, on grounds of technical complexity, trade secrecy or cybersecurity. Which data sets, which

variables and which weighting criteria the decision was based on may become invisible to the individuals who are the addressees of the decision (Yılmaz, 2020; Veale & Borgesius, 2021). This situation directly conflicts with the administration's obligation to give reasoned decisions. The principle of the reasoned decision is not merely a formal obligation of explanation; it requires that the factual and legal bases of the administrative act be set out in an intelligible manner (Gözler, 2020).

In the context of algorithmic administration, the statement of reasons cannot remain limited to the explanation of the final outcome alone. The explanation, even if only at a minimum level, of the fundamental criteria on which the decision rests, the types of data used and the logic of automated evaluation is imperative for the protection of individuals' right to an effective remedy. Otherwise, even if the right of access to the administrative courts exists in theory, it may become practically impossible for individuals to assert concretely the unlawfulness of the administrative act (Bayraktar, 2021; Yurtseven, 2021).

Another critical dimension of algorithmic administration with regard to the principle of the rule of law is the transformation of the administration's discretionary power. In classical administrative law, while discretionary power grants the administration a certain freedom of action, this freedom is limited by the principles of public interest, equality and proportionality. In algorithmic systems, by contrast, discretionary power is taken from the individual decision-maker and transferred to pre-coded rules and data-based modelling. Although this may appear on the surface to reduce arbitrariness, in reality it can lead to discretionary power being exercised in an indirect manner far removed from transparency (Bovens & Zouridis, 2002; Demir, 2020).

Risk scoring, automated classification and predictive analytics models in particular harbour the danger of concealing the administration's responsibility behind a discourse of "technical necessity" or "algorithmic neutrality". Yet the design, training and application of algorithms involve value-laden choices, and these choices do not eliminate administrative responsibility (Floridi vd., 2018; Şahin, 2022). For this reason, algorithmic administration must be treated as a form of governance that does not reduce the

administration's responsibility but, on the contrary, requires a more intensive accountability regime.

In this context, the risk-based regulatory approach adopted by the Bill on Artificial Intelligence makes it imperative that algorithmic administration be limited in a manner consistent with the principle of the rule of law. That algorithmic decisions not be accepted as absolute truths, that human oversight be ensured in a real and effective manner, and that these decisions be kept open to judicial review is not a normative choice but a constitutional necessity (Anayasa Mahkemesi, 2015). Otherwise, the risk will arise of the principle of the rule of law being eroded on grounds of administrative effectiveness and speed, and of the administration's discretionary power in practice escaping review.

For this reason, algorithmic administration should be evaluated not merely as a process of technical modernisation but as a field of normative transformation that redefines the transparency, accountability and legal responsibility of the administration. The established case law of the Turkish high courts on administrative discretion, proportionality and the principle of the reasoned decision offers a strong reference point as to how algorithmic administration can be reviewed within the framework of the rule of law. In this respect, the discussion of algorithmic administration constitutes the theoretical and normative ground of the analysis of Turkish judicial practice to be addressed in the next section.

5. Legal and Ethical Risks Concerning Artificial Intelligence Applications in Türkiye and Concrete Examples

In order to give concrete form to the legal framework concerning artificial intelligence applications in Türkiye, the aim has been to set out systematically, by means of the table below, the fundamental risks, the areas of liability and the relevant examples of application.

Table 3: Fundamental Risks, Areas of Liability and Examples of Application Concerning Artificial Intelligence Applications in Türkiye

Risk/Subject Heading	Explanation	Concrete Example of Application	Legal / Ethical Dimension	Source/Citation
Algorithmic Decision-Making	The use of AI in social assistance or public service decisions	The evaluation of social assistance applications by AI via e-Devlet	Risk of discrimination, unlawfulness of administrative decisions, principle of equality	Bayraktar, 2021; Yurtseven, 2021
Deepfake and Reputation	Fake images and videos produced with AI	The sharing of deepfake videos of celebrities or citizens on social media	Personality rights, violation of private life, constituting an offence under the Turkish Penal Code (TCK)	Öztürk, 2020; Sevinç, 2023
Autonomous Vehicles and Traffic	Driverless vehicles operating with AI	Autonomous vehicle accident tests in Istanbul and Ankara	Liability for damages under the Civil Code and the Code of Obligations, need for insurance law	Gless, Silverman & Weigend, 2016
Health Decisions	AI intervention in diagnosis and treatment processes	AI-supported radiology software making a misdiagnosis	Malpractice, violation of patient rights, incompatibility with the Health Services Law	Acar, 2018
Recruitment and Discrimination	CV screening and pre-selection with AI	AI-based recruitment systems discriminating on the basis of age, gender or education	Labour Law, principles of equal employment, ethical responsibility	Demirtaş, 2019
Data Privacy and Security	Data use by AI systems	The processing of personal health or financial data with large data sets	Law on the Protection of Personal Data, General Data Protection Regulation, personal data security, ethical data management	Anayasa Mahkemesi, 2015; Floridi vd., 2018

6. Assessment from the Standpoint of Turkish Judicial Practice

While direct and comprehensive case law specific to artificial intelligence is seen to be still limited in Turkish judicial practice, the established principles that the high courts have developed in the areas of digitalisation, automated decision-making processes, administrative discretion and interference with fundamental rights offer a strong normative framework that can be applied by analogy in resolving disputes arising from artificial intelligence. This situation also shows that artificial intelligence law will be a field that "evolves through case law".

6.1. Personality Rights and Digital Content in the Context of the Case Law of the Court of Cassation

In its decisions concerning the violation of personality rights in the digital environment, the Court of Cassation consistently accepts that internet publications may produce graver consequences because of the breadth and permanence of their sphere of impact. This approach is of direct importance for the evaluation of content generated by artificial intelligence.

Indeed, the 4th Civil Chamber of the Court of Cassation has stated that measures such as access blocking and content removal against digital content that violates personality rights may be legitimate, but that these interventions must comply with the principles of proportionality and necessity (Yargıtay 4. HD, 2017). This line of case law makes it imperative that interventions directed at deepfake content be evaluated not in an automatic and generalised manner but according to the particular features of the concrete case.

As also stated in academic studies, the fact that content has been generated by artificial intelligence does not eliminate liability; on the contrary, it raises the legal liability of the actors who put the content into circulation and enable its dissemination (Öztürk, 2020; Sevinç, 2023).

6.2. The Case Law of the Constitutional Court and Fundamental Rights

The decisions of the Constitutional Court concerning the protection of personal data, freedom of expression and the principle of equality draw the

constitutional limits of artificial intelligence regulations. The Court consistently maintains that interference with fundamental rights can be considered legitimate only where it is prescribed by law, proportionate and in conformity with the requirements of the democratic order of society (Anayasa Mahkemesi, 2015).

Within this framework, in the implementation of the Bill on Artificial Intelligence, it emerges as a constitutional necessity that the consequences produced for individuals, in particular by automated decision-making processes, be open to effective remedies. Otherwise, the risk will arise of indirect violations of rights being legitimised through technological instruments.

7. Conclusion and General Assessment

The Bill on Artificial Intelligence is an important step taken towards moving beyond the scattered and indirect normative framework of artificial intelligence law in Türkiye and establishing a systematic regulation specific to this field. At the same time, it is clear that the legal and institutional effect of the bill will be determined not merely by the existence of norms but by how these norms are given concrete form, how they are applied and how they are reviewed. For this reason, the success of the regulation is directly linked, as much as to its normative choices, to the secondary legislation, administrative practice and judicial case law that will accompany these choices.

First of all, the placing of the risk-based regulatory approach at the centre of the bill is apposite. However, for this approach to be effective, defining the risk categories merely through general concepts is not sufficient. The legislator must make the criteria concerning high-risk artificial intelligence applications clear and auditable through secondary legislation. Otherwise, the concept of risk carries the danger of drifting away from legal certainty and turning into an instrument that expands the administration's discretionary power and prepares the ground for arbitrary practices. In this context, tying risk classifications to transparent criteria and disclosing them regularly to the public should be regarded as a normative necessity.

Secondly, while the broad and flexible approach to the definition of the artificial intelligence system is advantageous in terms of adapting to

technological developments, it should be supported by limiting mechanisms in terms of legal foreseeability. Particularly in cases giving rise to sanctions, access blocking and liability, technical guidelines and binding administrative guidance should be prepared so that vague definitions do not create a chilling effect on individuals. At this point, the task falling to the legislator is to strengthen certainty at the normative level rather than to refer uncertainty to the judiciary.

With regard to the liability regime, the non-recognition of artificial intelligence as an independent legal subject is appropriate. However, if the limits of user and developer liability are left to general principles alone, serious problems of proof and imputability will arise in practice. For this reason, it is proposed that differentiated liability models be explicitly regulated with respect to phases such as the design stage, data selection, the training process and the supervision of use. Particularly in high-risk systems, bearing in mind that the classical approach, which places the burden of proving fault entirely on the injured party, will prove insufficient, models of a reversed burden of proof or objectivised liability need to be opened up for discussion.

In the area of deepfake content and algorithmic content generation, meanwhile, it is imperative that the regulation adhere strictly to the principle of proportionality. The existence of rapid intervention mechanisms does not by itself confer legitimacy from the standpoint of the democratic order of society. These mechanisms must without fail be operated on the basis of reasoned and individualised decisions open to judicial review. Otherwise, the aim of combating risks arising from artificial intelligence may turn into systematic and automatic restrictions on freedom of expression. The approach that should be proposed normatively is that the principle of "subsequent and proportionate intervention" be taken as the basis, rather than "preventive censorship".

With regard to discriminatory data sets and algorithmic inequality, while the orientation of the bill is correct from a constitutional standpoint, independent audit mechanisms are needed for it to be effective in practice. It should be ensured that the training data of artificial intelligence systems can be audited not only by the administration but, where necessary, by independent and expert boards. This audit should, beyond being a technical

conformity check, involve a normative assessment on the basis of the principle of equality and the prohibition of discrimination.

Finally, the principles of transparency and explainability need to be made an explicit obligation, particularly with regard to algorithmic systems used by public administration. For the administration to keep its decision-making processes closed on the grounds of "technical complexity" is incompatible with the principle of the rule of law. For this reason, it should be normatively guaranteed that individuals can access the fundamental reasons for the algorithmic decisions taken about them and can make use of effective remedies against these decisions.

While the Bill on Artificial Intelligence is necessary and apposite for the development of artificial intelligence law in Türkiye, it also offers a framework whose content needs to be supported. The real function of the regulation lies in its ability to create a dynamic field of law continuously reproduced through administrative practice, judicial case law and academic scrutiny, rather than in being a static body of norms. In this context, artificial intelligence law should be treated, for Türkiye, not merely as a field of technological regulation but, in future studies, as a field of normative choice concerning how the principle of the rule of law is to be protected in the digital age.

References

Anayasa Mahkemesi. (2015). *E. 2014/149, K. 2015/45*. *Anayasa Mahkemesi Kararlar Dergisi.*

Acar, M. (2018). *Otonom sistemlerin hukuki sorumluluğu* (Yüksek lisans tezi, Gazi Üniversitesi, Sosyal Bilimler Enstitüsü). Gazi Üniversitesi.

Akıncı, A. (2021). Yapay zekâ ve hukuk: Temel kavramlar ve sorumluluk tartışmaları. *Ankara Üniversitesi Hukuk Fakültesi Dergisi, 70*(2), 345–378.

Arslan, Z. (2022). *Yapay zekâ uygulamalarının temel hak ve özgürlükler üzerindeki etkisi* (Doktora tezi, Ankara Hacı Bayram Veli Üniversitesi, Lisansüstü Eğitim Enstitüsü). Ankara Hacı Bayram Veli Üniversitesi.

Bayraktar, H. (2021). Algoritmik idare ve hukuk devleti ilkesi. *İdare Hukuku Dergisi, 35*(1), 55–82.

Birleşmiş Milletler. (2024, 21 Mart). *Yapay zekâ ve insan haklarına ilişkin Genel Kurul kararı* [Çok taraflı karar]. Birleşmiş Milletler Genel Kurulu.

Demir, B. (2020). Dijital çağda hukuk devleti ve yapay zekâ. *Hukuk Araştırmaları Dergisi, 12*(3), 101–130.

Demirtaş, F. (2019). İş hayatında yapay zekâ ve ayrımcılık: Hukuki ve etik boyutlar. *Çalışma ve Toplum Dergisi, 24*(2), 45–68.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines, 28*(4), 689–707.

Gless, S., Silverman, B., & Weigend, T. (2016). Autonomous systems and liability: Legal perspectives on self-learning machines. *Journal of Technology Law, 22*(1), 15–38.

Gözler, K. (2020). Yapay zekâ ve Türk hukuku: Düzenleme ihtiyaçları. *Hukuk ve Teknoloji Dergisi, 7*(2), 33–56.

Kara, M. (2019). Yapay zekâ ve ceza hukuku: Sorumluluk ve kusur. *Ceza Hukuku Araştırmaları Dergisi, 11*(1), 77–102.

Koç, E. (2021). *Yapay zekâ ve ceza hukuku sorumluluğu* (Doktora tezi, İstanbul Üniversitesi, Hukuk Fakültesi). İstanbul Üniversitesi.

Özgenç, H. (2020). Yapay zekâ sistemlerinin hukuki statüsü ve sorumluluk rejimi. *Ankara Barosu Dergisi, 78*(4), 215–240.

Öztürk, H. (2020). Dijital içerik ve deepfake: Hukuki sorumluluk. *Bilişim Hukuku Dergisi, 5*(1), 45–67.

Sevinç, A. (2023). Yapay zekâ kaynaklı içerik üretimi ve ifade özgürlüğü. *İnternet ve Hukuk Dergisi, 9*(1), 12–38.

Şahin, T. (2022). Ayrımcı algoritmalar ve veri güvenliği: Hukuki çerçeve. *Bilişim Hukuku Araştırmaları Dergisi, 10*(2), 59–81.

TBMM. (2025). *Yapay zekâya ilişkin kanun teklifi: 07 Kasım 2025* [Kanun teklifi]. Türkiye Büyük Millet Meclisi.

Yargıtay 4. Hukuk Dairesi. (2017). Erişim engeli ve içerik kaldırma kararları. *Yargıtay Kararları Dergisi, 9*, 112–125.

Beyond One-Size-Fits-All: Artificial Intelligence in Foreign Language Education for Gifted Learners

Ayşegül İşlekeller Bozca[1]

1. Introduction

It is natural and easy to accept the truth that "one size does not fit all," yet in practice this truth is often overlooked. Del Siegle notes that what is obvious for a shoe or a garment should hold equally for education. Students arrive with different experiences and learn at different rates; indeed, a single fifth-grade classroom may span an achievement range of five to seven grade levels (Siegle, 2024). Given its capacity to respond to this diversity, artificial intelligence (AI) is regarded as a technology that could reshape education, particularly through its affordances for personalized learning and instructional differentiation (Guilbault et al., 2025). Increasingly common in education, AI tools offer new possibilities for instructional processes such as lesson design, differentiation, and feedback (Guilbault et al., 2025). Yet how genuinely inclusive AI's personalization affordances are must also be questioned, because personalization is meaningful only to the extent that it reaches all students with differing needs. Indeed, when teachers differentiate, they tend to turn to the students who need the most support (Siegle, 2024). This tendency can lead differentiation to be conceived largely in terms of students who experience learning difficulties or who require additional support to reach grade level. By contrast, the needs of gifted students—those who can progress beyond grade level and who require different learning opportunities in terms of pace, depth, complexity,

[1] University of İstanbul-Cerrahpaşa, Faculty of Education, Department of Special Education, Department of Education of People with Hearing Disability ORCID: 0000-0003-2912-5379, *islekellerbozca@iuc.edu.tr,* Türkiye

or intellectual challenge—may remain less visible (Siegle, 2024). For this reason, rather than fully resolving the problem of differentiation, AI has brought an important question back to the fore: for whom is differentiation actually carried out?

The answer to this question reveals the paradox at the heart of this chapter: students who progress quickly in the foreign language classroom, who gravitate toward more complex learning experiences, and who are ready to move beyond the existing content may at times encounter insufficient cognitive challenge. Standard foreign language instruction structured around the average student can remain limited in terms of pace, depth, and complexity, particularly for students who master content quickly. In assessing this picture for gifted students, the diversity within the group itself must be kept in mind. Although general cognitive ability and foreign language aptitude may be related, they do not represent the same construct. Even so, some gifted students—those with advanced verbal skills or strong foreign language aptitude—may show marked strengths in vocabulary knowledge and language learning performance. Components of foreign language aptitude such as phonological processing, the analysis of linguistic patterns, and lexical learning are likewise associated with individual differences in second language development (Bain et al., 2010; Serafeim & Alexiou, 2020; Sun et al., 2026). Yet the rationale for differentiation for gifted students is not confined to linguistic strengths. Their needs to progress more rapidly, to work with deeper and more complex content, to pursue strong interests, and to engage with intellectually demanding tasks also shape how foreign language instruction should be structured (Prodanovska-Poposka & Todorova, 2020; Siegle, 2024). Learning environments that respond to these needs support students' engagement and development, whereas insufficient cognitive challenge can be associated with a loss of motivation and disengagement from learning (Prodanovska-Poposka & Todorova, 2020). The central issue, therefore, is not that gifted students maintain their current levels of achievement, but that they gain access to learning opportunities involving the pace, depth, autonomy, and cognitive challenge that can develop their potential (García-López et al., 2025).

At this point, the significance of AI for the foreign language education of gifted students arises not merely from its ability to personalize instruction,

but from its potential to organize the learning experience more flexibly according to the student's pace, readiness, interests, and cognitive needs. Systems based on generative AI (GenAI), machine learning, and natural language processing offer tools that can partly move beyond the limits of standardized instruction through affordances such as adapting the difficulty level of tasks, providing students with alternative content pathways, delivering immediate feedback, and supporting independent progress (García-López et al., 2025; Guilbault et al., 2025; Siegle, 2023). These affordances appear pedagogically meaningful particularly for gifted students, who may need to progress more rapidly, work with deeper and more complex content, engage in interest-based learning, and exercise a higher degree of autonomy.

This chapter is positioned at the intersection of the rapidly expanding uses of AI in foreign language education and the differentiated learning needs of gifted students. Rather than offering a technology-centered review of tools, it undertakes a conceptual synthesis that evaluates AI's pedagogical affordances through established theoretical principles in gifted education. To this end, Renzulli's approach to gifted behavior and enrichment and Siegle's framework of acceleration, depth, complexity, and interest-based learning serve as its foundation.

At this point, the significance of AI for the foreign language education of gifted students arises not merely from its ability to personalize instruction, but from its potential to organize the learning experience more flexibly according to the student's pace, readiness, interests, and cognitive needs. Systems based on generative AI (GenAI), machine learning, and natural language processing offer tools that can partly move beyond the limits of standardized instruction through affordances such as adapting the difficulty level of tasks, providing students with alternative content pathways, delivering immediate feedback, and supporting independent progress (García-López et al., 2025; Guilbault et al., 2025; Siegle, 2023). These affordances appear pedagogically meaningful particularly for gifted students, who may need to progress more rapidly, work with deeper and more complex content, engage in interest-based learning, and exercise a higher degree of autonomy.

This chapter is positioned at the intersection of the rapidly expanding uses of AI in foreign language education and the differentiated learning needs

of gifted students. Rather than offering a technology-centered review of tools, it undertakes a conceptual synthesis that evaluates AI's pedagogical affordances through established theoretical principles in gifted education. To this end, Renzulli's approach to gifted behavior and enrichment and Siegle's framework of acceleration, depth, complexity, and interest-based learning serve as its foundation.

2. Theories of Giftedness and AI Integration

Gifted education has long moved beyond answering the question "Who is gifted?" with a single score from an intelligence test. Today, giftedness is not a fixed label but a form of behavior that emerges in students under certain conditions (Renzulli, 2016). Understanding the role of AI technologies in the foreign language education of gifted students likewise cannot be limited to examining the technical capacities of these tools. Technology can acquire educational value in the field of gifted education only when it is built upon theories of gifted education (Renzulli, 2012). In this regard, Joseph Renzulli's Three-Ring Conception and Schoolwide Enrichment Model (SEM), together with Del Siegle's Three-Legged Stool of talent development, provide a well-suited theoretical foundation (Renzulli, 1978/2021, 2005b, 2012; Renzulli & Reis, 1997, 2014; Siegle, 2024). These theoretical models allow us to ground the personalizing, enriching, and accelerating capacity of AI in foreign language acquisition (García-López et al., 2025; Siegle, 2024).

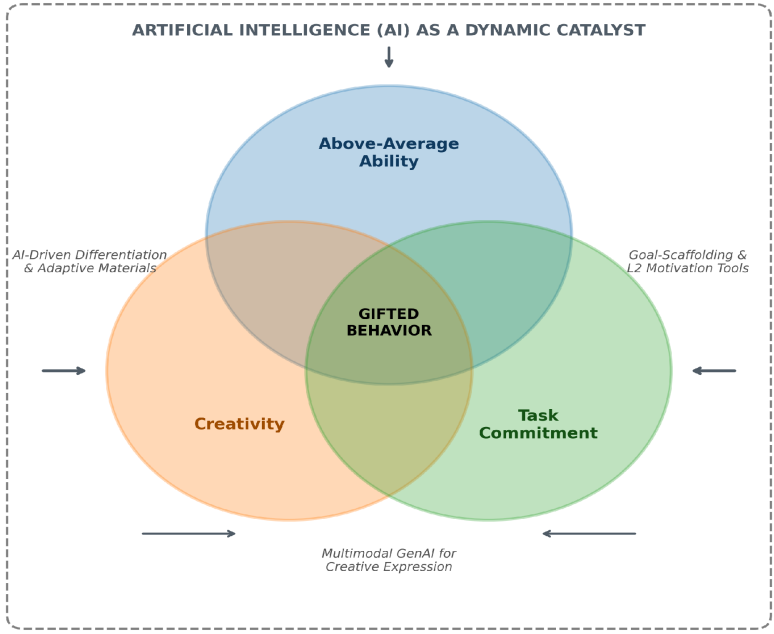
2.1. Renzulli's Three-Ring Conception and AI

The Three-Ring Conception developed by Joseph Renzulli treats giftedness as a form of behavior that emerges in particular students under particular conditions (Renzulli, 1978/2021, 2005b, 2016). For this behavior to arise, three clusters of traits—above-average ability, task commitment, and creativity—must interact; no single cluster "produces giftedness" on its own, and what is decisive is the interaction among them (Renzulli, 2016). This understanding has two important implications. First, because giftedness is a developable capacity rather than a static one, the classroom's task is as much to "develop" it as to "find" it. Second, Renzulli's distinction between "schoolhouse giftedness," which manifests in academic achievement, and

"creative-productive giftedness," which is oriented toward generating original work (Renzulli, 2016, p. 62), reminds us that what standardized tests measure does not always coincide with what we actually seek to develop in the classroom.

According to the conception, the intersection of the rings—the area where creative-productive output emerges—is the domain of high ability and gifted performance (Renzulli, 1978/2021, 2005b). The pedagogical counterpart of this definition takes shape in three mutually reinforcing lines of intervention. The Schoolwide Enrichment Model (SEM), in which the conception is put into practice, consists of general exploratory experiences that introduce students to a broad range of topics (Type I), group training activities that develop thinking and research skills (Type II), and, finally, individual or small-group work devoted to the original investigation of a real problem (Type III); this last is the model's most advanced form of enrichment (Renzulli, 2012).

Figure 1 Renzulli's Three-Ring Conception of Giftedness with AI as a Catalyst



Note. This figure was conceptualized and developed by the author, integrating artificial intelligence (AI) as a dynamic external catalyst within Joseph Renzulli's Three-Ring Conception of Giftedness (Renzulli, 2012, 2016). AI functions as an overarching catalyst that enhances and scaffolds above-average ability, creativity, and task commitment in foreign language education for gifted learners.

In other words, AI technologies can function as a "catalyst" that nourishes these three rings and the enrichment process, triggering the dynamic interaction among them (García-López et al., 2025; Siegle, 2024).

a. **The Above-Average Ability Ring and AI.** The fixed pace and level of traditional foreign language classrooms can create an "instructional ceiling," particularly for gifted students whose current language proficiency is above grade level or who display advanced verbal skills and high language aptitude (Siegle, 2024). The core problem here is not that all gifted students perform exceptionally in a foreign language, but that standardized instruction does not respond adequately to differences in readiness and learning pace among students. AI-supported tools can make it easier to differentiate the linguistic complexity of texts, the difficulty level of tasks, and the support provided according to the student's current performance (García-López et al., 2025; Siegle, 2024). In this way, students with advanced language proficiency can work with more complex lexical, syntactic, and discursive structures, while other gifted students can encounter tasks that are suited to their own readiness levels yet remain cognitively challenging.

b. **The Creativity Ring and AI.** Renzulli defines creativity as curiosity, originality, and the capacity to challenge conventional patterns (Renzulli, 1978/2021, 2005b). GenAI models can enable students to transform words into visual, auditory, and textual forms (e.g., text-generating ChatGPT and text-to-image DALL·E) (Siegle, 2023). In learning a language, students do not merely consume ready-made texts; using AI as a creative partner, they produce original digital content in the target language (Acar, 2023; Siegle, 2023).

c. **The Task Commitment Ring and AI.** This ring represents perseverance, determination, focused motivation, and the energy to work on a problem over an extended period (Renzulli, 1978/2021, 2005b). Renzulli argues that this affective dimension is the most critical and situational component of gifted behavior (Renzulli, 1978/2021, 2005b). AI-based intelligent tutoring assistants can contribute to sustaining task commitment and intrinsic second-language (L2) motivation by providing feedback tailored to the student's personal interests and by breaking tasks into manageable sub-goals (e.g., with tools such as Goblin.tools) (Siegle, 2024; Wei, 2023).

When a student interacts with AI, they encounter original problems that spark their interest and thereby find dynamic situations—Type I and Type II enrichment opportunities—in which to display their creativity and task commitment (Renzulli, 2005b, 2012; Renzulli & Reis, 1997). This integrated view presents what the gifted student needs not as a single list but as a system of mutually interacting elements. The needs of this system can be gathered under three headings, and Siegle's framework of the "three legs of talent development" supports this consolidation.

2.2. Del Siegle's Three-Legged Stool Model and AI Integration

Del Siegle conceptualizes the core service models in gifted education as the "three-legged stool" of talent development (Siegle, 2024). Each of these legs plays a critical load-bearing role in supporting and sustaining a gifted student's intellectual development; the absence of any one leg causes the stool to collapse (Siegle, 2024). Siegle (2024) argues that AI technologies hold the potential to support each of these three legs.

a. Leg 1: Academic Acceleration and Curriculum Compacting

Gifted language students have a learning pace much faster than the class average (Reis et al., 2016; Siegle, 2024). Curriculum compacting is a systematic differentiation strategy that identifies the topics and skills these students have already mastered and thereby streamlines the curriculum for them, freeing time for more demanding work (Reis et al., 2016; Tomlinson, 2014). Renzulli and Reis recommend that students who score at least 80–85% on pre-assessments be exempted from routine repetition within that unit (Reis et al., 2016; Siegle, 2024).

The most laborious aspect of compacting is continually determining which student already knows what, which is difficult in practice for a single teacher in a crowded classroom. It is precisely here that AI's capacity to offer individualized learning plans, immediate feedback on writing, and complex questions and challenges (García-López et al., 2025) can come into play, providing an adaptive mechanism that quickly diagnoses mastered content and moves the student to the next level.

Consider a gifted student of ours who has grasped the grammatical structure to be covered that day well before most of the class. In the traditional flow,

this student is obliged to listen once again to material they already know, and their attention gradually drifts. With AI support, the student's readiness can be analyzed in a short time and tasks appropriate to the areas in which their understanding is incomplete can be assigned.

It is important to read the gain here correctly. What is gained is time; but time is not an end in itself. The real question is what the space opened up for this student will be filled with, because moving a gifted student along merely "faster" will sooner or later leave them bored once again. This brings us to the second and more decisive need: depth. Some teacher-facing AI tools can enable the rapid preparation of pre-assessment and placement materials (e.g., MagicSchool AI) (García-López et al., 2025; Siegle, 2024). Intelligent tutors that adopt a Socratic approach can adapt to the student's level; rather than giving answers directly, some examples guide the student to think and keep a progress record for the teacher (e.g., Khanmigo) (Siegle, 2024). The product names cited throughout this chapter are examples current at the time of writing; given the field's rapid change, the emphasis is not on specific products but on the functions they represent.

b. Leg 2: Inductive Teaching, Depth, and Complexity

Pace alone is not a solution; moving a student more quickly through the same kinds of tasks only bores the gifted student sooner. What the gifted student truly needs is qualitatively different, deeper, and more challenging content. Prodanovska-Poposka and Todorova emphasize that advanced-level content is essential for these students because they progress rapidly (Prodanovska-Poposka & Todorova, 2020), and they make this concrete for each language skill in turn. For speaking and thinking, they note that higher-order question stems and provocative "what if?" questions push the student toward deeper thinking (Prodanovska-Poposka & Todorova, 2020). For reading, they recommend that advanced readers be offered texts above grade level that contain more demanding literary figures—such as allegory, metaphor, and parable—and that call for creative and critical thinking from the reader (Winebrenner, 2016, as cited in Prodanovska-Poposka & Todorova, 2020).

This leg aims to keep the student from consuming information superficially and to direct them toward higher-order analytical processes. In the Kaplan

framework conveyed by Siegle, depth refers to examining a topic's internal language, details, patterns, rules, ethical dimensions, and as-yet-unanswered questions, whereas complexity refers to approaching a topic through its change over time, different points of view, and relationships with other disciplines (Kaplan, 2009, as cited in Siegle, 2024). The inductive approach also nourishes this depth, because gifted students tend to seek an understanding deeper than surface-level knowledge, and this inquiry-based method fosters precisely such understanding (Siegle, 2024).

GenAI and large language models (LLMs) can address this need directly. Through interactive exchanges with tools such as ChatGPT or Perplexity, students can delve more deeply into a topic or explore interdisciplinary connections (Siegle, 2024). This use can take two concrete forms. The first is interdisciplinary, multiperspective inquiry that foregrounds multiple viewpoints, change over time, and cross-disciplinary links; a student may, for example, investigate in depth the historical evolution of a concept or a complex topic that cuts across several fields. The second is inductive analysis: AI-powered search engines that cite their sources can guide the student—by indicating, alongside their summaries, the sources from which the information is drawn—to verify information, subject it to critical scrutiny, and reach their own inferences from specific examples (Siegle, 2024).

The contribution of AI in this area arises from its generative nature. GenAI does not merely deliver ready-made content; it can also produce new content and thereby open up space for students to develop their creativity (García-López et al., 2025). AI transformers can, by presenting students with different viewpoints and scenarios, push them to enhance their creativity and critical thinking, and can even facilitate collaboration by acting as a kind of virtual partner (García-López et al., 2025). For the gifted student, this means being able to move beyond exercises with a single "correct answer" toward open-ended, authentic tasks in which multiple perspectives are debated.

The one who benefits most from this tool is often the gifted student. Thanks to their strong analytical skills, creativity, and ability to connect different domains, these students can be said to be well suited to prompt engineering, the pivotal skill of AI interaction (Siegle, 2025). Their capacity for interdisciplinary thinking allows them to construct more contextual and nuanced

prompts, and their ability to break complex problems into parts enables them to obtain higher-quality outputs from AI (Siegle, 2025). Siegle proposes strategies for structuring this interaction, such as the CRAFT model, prompt chaining, and the flipped interaction pattern (Siegle, 2025). Moreover, the gifted student's critical and creative thinking serves not only in constructing better prompts but also in evaluating the quality of the AI's response (Siegle, 2025). Prompt engineering thus becomes not an ordinary "search" skill but a discipline of higher-order questioning; and the way conversational interaction lends itself to deepening through follow-up questions (Siegle, 2023) can help sustain this discipline. Case-based studies describing how these affordances might be used at the secondary level likewise address differentiation, personalization, and independent research together (Guilbault et al., 2025).

Let us therefore take our earlier scenario one step further: rather than moving our student—who grasped the day's grammatical structure well before most of the class—on to a new topic, let us guide them to go deeper within the same structure. Instead of stating the rule directly, the teacher used AI tools to prepare a list of examples in which the structure appears in different sentences and asked the student to discover the rule themselves. The student examined the examples one by one, noticed the similarities among them, and inferred for themselves the situations in which the structure is used. The rule was thus not handed to the student ready-made; they discovered it themselves, starting from the examples. This is inductive teaching.

The teacher then sought to have the student think about the same topic more deeply and from more angles. Again drawing on AI, the teacher posed questions: How does this structure change in everyday speech versus formal writing? Has its use shifted over time? How is the same meaning conveyed in another language you know? These questions carried the student beyond a single correct answer, prompting them to consider the topic from different perspectives, in comparison with other languages, and together with its everyday usage. In this way, the student engaged with the same structure both more deeply and in a more complex manner. This is depth and complexity.

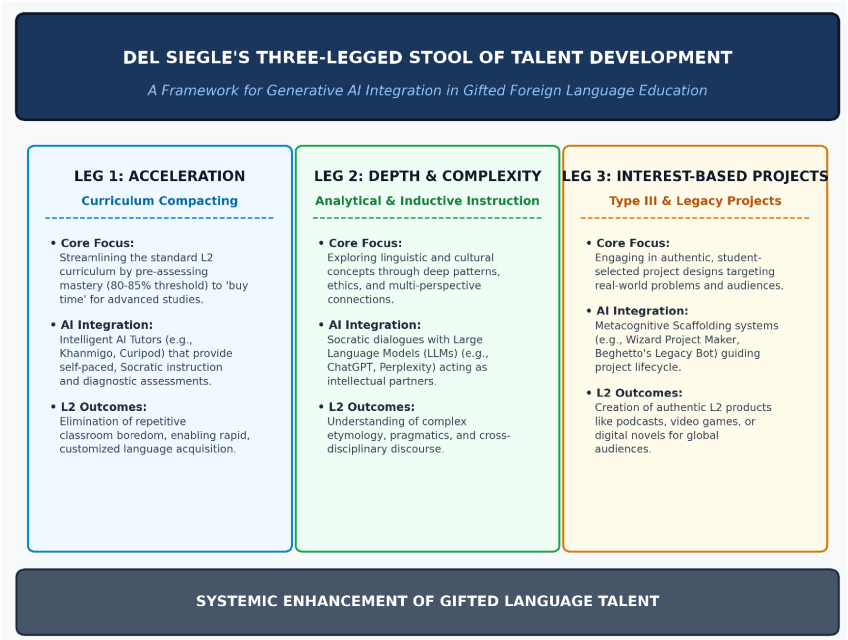
c. Leg 3: Interest-Based Activities and Original Products (Type III Enrichment)

The final leg of the stool offers students the opportunity to design original products and interest-based work in which they can generate solutions to real-world problems along the lines of their own interests and share these products with a genuine audience (Siegle, 2024). This approach converges with Type III enrichment—the most advanced level of Renzulli's Enrichment Triad, namely the original investigation of a real problem at the individual or small-group level (Renzulli, 2012)—and it also resembles Beghetto's Legacy Projects approach, in which students aim to make a lasting contribution to their community (Beghetto, 2017, as cited in Siegle, 2024).

In this leg, AI can support the student's self-direction, enabling them to progress along the lines of their own interests and independently of the teacher's pace (García-López et al., 2025; Siegle, 2024). Tools that guide the student in carrying out such projects also exist; for example, the project-development component of the Renzulli Learning System (the Wizard Project Maker) (Renzulli & Reis, 2007) or the bot Beghetto designed on ChatGPT to guide the student step by step through the Legacy Project process (Siegle, 2024) can support the student in planning and carrying out their project. In producing an original work in the target language for a genuine audience of readers or listeners, the student does not merely learn the language but uses it for a real purpose. These products—considered solely to make the concept concrete, and not as a source-based finding—may take various forms, such as a digital story, a podcast episode, or a small digital game in the target language.

The self-management and graduated support (scaffolding) that such original-product processes require are addressed in detail in Section 4.2.

Figure 2 Del Siegle's Three-Legged Stool of Talent Development and AI Integration



Note. Developed by the author, based on Siegle's (2024) Three-Legged Stool of Talent Development model. Each leg represents a critical pedagogical pillar supporting the intellectual and linguistic development of gifted language learners, illustrating how artificial intelligence may scaffold, accelerate, and enrich these dimensions.

Table 1: Comparative Synthesis of Renzulli's Three-Ring Conception and Siegle's Three-Legged Stool Models Integrated with AI in Foreign Language Education

Theoretical Model & Core Components	Pedagogical Focus (Traditional Focus)	AI Integration & Digital Tools (AI Integration)	L2 Outcomes & Linguistic Deliverables (L2 Outcomes)
Renzulli: Above-Average Ability *(Above-Average Ability)*	• Supporting the student's rapid cognitive processing, phonetic sensitivity, and advanced verbal reasoning.	• **Adaptive Learning Systems & NLP:** Dynamic level diagnostics and lexically scaffolded/level-adjusted reading texts.	• Raising the traditional "instructional ceiling."• Support for acquiring advanced, domain-specific academic vocabulary.

Renzulli: Creativity *(Creativity)*	• Triggering curiosity, originality, fluency of ideas, and the capacity to challenge conventional paradigms.	• **Generative Writing & Multimodal AI:** Interactive dialogue with LLMs (ChatGPT/Gemini) and text-to-media rendering tools (DALL-E, Synthesia).	• Transition from mechanical and passive language consumption to authentic, multimodal digital content creation.
Renzulli: Task Commitment *(Task Commitment)*	• Sustaining intrinsic motivation, goal-directedness, persistence, and grit during demanding tasks.	• **Intelligent Process Management:** Segmenting complex, long-term tasks into highly manageable sub-goals (e.g., Goblin.tools).	• Sustained L2 motivation and prolonged focus/grit during highly demanding linguistic and communicative tasks.
Siegle Leg 1: Academic Acceleration *(Acceleration & Compacting)*	• Streamlining already mastered curriculum (requiring 80-85% pre-assessment mastery) to buy time for challenging enrichment.	• **Socratic AI Tutors & Diagnostics:** Automated diagnostic testing via MagicSchool AI/Curipod, and customized pacing with Socratic AI tutors (e.g., Khanmigo).	• Transcending the boundaries of linear, repetitive standard curricula; independent, self-paced progress.
Siegle Leg 2: Depth & Complexity *(Depth & Complexity)*	• Investigating rules, patterns, historical/cultural context, ethics, and interdisciplinary connections of language.	• **Large Language Models & AI Search:** Dialogic search, etymological inquiry, and cross-cultural discourse analysis using ChatGPT or Perplexity AI.	• Inductive and structural analysis of language; critical evaluation of complex linguistic, academic, and cultural discourses.
Siegle Leg 3: Interest-Based Creative Products *(Type III & Legacy Projects)*	• Formulating creative solutions to real-world problems to create a legacy and impact a real audience.	• **Metacognitive Project Wizards:** Guided research and project workflow management using Renzulli's Wizard Project Maker and Beghetto's Legacy Project Bot.	• Opportunities to produce and share authentic L2 products (e.g., podcasts, digital stories, or small games in the target language).

Note. This table presents a theoretical synthesis of the gifted education frameworks of Joseph Renzulli (1978/2021, 2005b, 2016) and Del Siegle (2024), developed by the author to illustrate how generative artificial intelligence may be integrated into foreign language education for gifted learners

3. AI-Based Differentiation of Language Skills for Gifted Students

For gifted students, differentiation in foreign language instruction does not mean simply offering more advanced linguistic content. A student's proficiency level in the target language must be distinguished from the level of challenge they need cognitively; a student may require tasks that demand greater depth, complexity, choice, and original production even before attaining advanced language proficiency. For this reason, the four core skills are examined below not in terms of general AI applications but in terms of the acceleration, depth and complexity, interest-based learning, and creative-productive activities discussed in the previous section.

Reading. Gifted readers tend to acquire target-language vocabulary and grammar early, to decode complex figures such as allegory and metaphor rapidly, and to use reading as a means of inquiring into interdisciplinary connections (Prodanovska-Poposka & Todorova, 2020). AI can differentiate reading input in two directions: on the one hand, it can raise a text to a level well above the class grade, with more complex syntax and more advanced vocabulary (Guilbault et al., 2025; Siegle, 2024); on the other, AI-powered search engines that cite their sources can enable students to read deeply in the target language on advanced topics not covered in the traditional curriculum and to verify information by following the sources provided (Siegle, 2024). The point specific to gifted learners here is not that the text is made easier for the student, but that it can be deepened beyond the class grade level; differentiation responds to the need for depth and complexity when it is framed not as "more accessible text" but as "more demanding text."

Writing. Writing is a productive skill that requires organizing one's cognitive processes at the highest level; gifted students are eager to express themselves in original ways in the target language, yet the mechanical corrections involved in the writing process can dampen their motivation (Prodanovska-Poposka & Todorova, 2020; Siegle, 2023). Here, GenAI models can be used as a brainstorming and drafting partner; when operated as a partner that asks guiding questions rather than supplying ready-made text, they can help the student structure their argument and draw on their metacognitive skills (García-López et al., 2025; Guilbault et al., 2025; Siegle, 2025). Multimodal generative tools, in turn, can open up space for creative expression by transforming a description

the student has written in the target language into a visual or auditory output (García-López et al., 2025; Siegle, 2023). Given the gifted student's creative-productive capacity, the value here lies in allowing an original voice to develop; the same tool, however, also carries the risk of homogenizing originality, a risk addressed in Section 6.

Listening. Gifted students possess a heightened phonetic sensitivity, yet the standard listening exercises in traditional classrooms often fall below their capacity to analyze complex discourse (Prodanovska-Poposka & Todorova, 2020; Siegle, 2024). AI can move listening beyond a static exercise and tie it to the student's interests: personalized listening materials can be generated in the target language on a topic the student is passionate about, while applications that bring historical figures to life can enable the student to interact in the target language with people from the past and to analyze their responses (García-López et al., 2025; Guilbault et al., 2025; Siegle, 2023). In this way, listening is linked to the student's interest-based deepening, and the input can be drawn up to the level of the student's intellectual curiosity.

Speaking. Speaking is the area of greatest anxiety for the gifted student, who may have perfectionist tendencies; the tension of not yet being able to express their complex thoughts fully in the target language can push them into silence (Prodanovska-Poposka & Todorova, 2020; Zhang et al., 2025). Chat-based AI tools can provide the student with an accessible and patient conversation-practice partner, while bots designed around inquiry can engage the student in scientific or philosophical questioning in the target language and offer immediate feedback (indeed, one such bot was used in a study conducted with gifted primary school students; Chang et al., 2023; García-López et al., 2025). Its affective counterpart—namely, reduced anxiety and increased willingness to communicate—is addressed in detail in Section 4.1.

4. Affective and Psychosocial Dynamics in Language Instruction

4.1. AI, Anxiety, and Willingness to Communicate

Learning a foreign language is not merely a cognitive process; whether a student actually uses the language depends largely on affective factors such as motivation, anxiety, and self-confidence. This dimension is especially

important for the gifted student, because these students have, alongside their cognitive needs, distinct social and emotional needs that must be recognized (García-López et al., 2025; Guilbault et al., 2025). It is precisely here that AI enters as a learning partner capable of influencing the affective climate of the learning environment (García-López et al., 2025).

Foremost among these affective factors is foreign language classroom anxiety: the tension and unease a student feels when using the target language, especially when speaking, is one of the most salient obstacles to using the language effectively. AI offers a notable affordance here; in AI-supported environments, students' AI literacy has been found to be associated with lower classroom anxiety (Zhang et al., 2025). A likely reason is that AI can provide a practice space in which the social cost of error is low and there is no anxiety about being judged; the student can try and err without feeling peer pressure. For the gifted student, this means a safer footing on which they can set aside their perfectionist anxieties and experiment.

The most concrete outcome of this affective climate is willingness to communicate (WTC): the student's readiness to initiate communication in the target language. WTC depends not on a single factor but on many, including the learning environment, the level of anxiety, and self-efficacy (Zhang et al., 2025). In AI-supported environments, these factors become interconnected: a student's AI literacy is associated with higher self-efficacy and lower anxiety, and these two emerge as mediating factors that increase WTC (Zhang et al., 2025). In other words, an environment that lowers anxiety and fosters self-confidence also raises the student's desire to use the language; for the gifted student, this means being able to carry their complex thoughts into the target language rather than remaining silent.

4.2. AI and Self-Regulated Learning

One of the most critical determinants of lasting success in a foreign language is the student's taking responsibility for their own learning process—that is, becoming an autonomous learner. This autonomy is closely related, at the cognitive level, to self-regulated learning (SRL) and, at the affective level, to second-language motivation (Wei, 2023). Self-regulated learning is a multidimensional construct encompassing processes such as setting goals, selecting strategies, monitoring one's own progress, and evaluating

one's performance, and it is strengthened in environments that afford students opportunities for choice (Housand & Reis, 2008).

This capacity carries particular importance for the gifted student and brings with it a subtle paradox. Because these students have a very fast learning pace, it can be difficult for them to develop independent-study skills within the confines of the standard curriculum; some gifted students, who learn everything at school with ease, may not sufficiently develop self-regulation skills such as exerting effort and planning, and this can lead in later years to underachievement (Housand & Reis, 2008). It is precisely here that AI can open up a self-regulation space in which the student can progress at their own pace, set personalized goals, and monitor themselves through immediate feedback (García-López et al., 2025; Wei, 2023).

The established principle that explains how this support works is scaffolding: an adult or a guide temporarily takes on the elements of a task that exceed the student's current capacity and gradually withdraws this support as the student becomes more competent (Wood et al., 1976). In the Vygotskian social-constructivist framework on which Wei draws, learning is at first an externally regulated activity that, over time, is internalized and becomes self-regulated; AI-based language platforms can construct this graduated support in a digital environment, offering hints and explanations when the student struggles and withdrawing support once mastery is achieved, leaving the student to grapple with higher-order tasks (Wei, 2023). In this process, the teacher can make metacognitive reflection visible; for example, an AI interaction that, after a task, poses questions to the student such as "Which linguistic structure challenged you most in this work, and what strategy did you use to overcome it?" can help the student analyze their own strategies (Guilbault et al., 2025).

5. The Teacher's Role and Task Design for the Gifted Language Student

The affordances outlined so far come to fruition in the classroom only through deliberate pedagogical design and the teacher's guidance. The entry of AI into the classroom transforms the teacher's traditional role. The teacher moves from being the sole source of knowledge to being a guide

who designs and manages the student's interaction with AI (Guilbault et al., 2025; Siegle, 2023). The role of the teacher who works with gifted students can be gathered under three categories: (1) serving as the student's mentor and supporter, (2) possessing AI literacy, and (3) managing the relationship between AI and their own judgment in assessment (García-López et al., 2025).

5.1. From Knowledge Provider to Facilitator of "Virtual Mentorship"

One of the greatest obstacles in gifted students' foreign language learning is that, owing to time constraints and curricular pressure, teachers cannot prepare materials differentiated to each student's interests and learning pace (Reis et al., 2016). By removing this constraint, LLMs and GenAI tools can make it possible to approach the ideal—articulated by Renzulli and Reis (2007)—of "a dozen teaching assistants" always at the teacher's side in the classroom. By positioning AI as a set of digital assistants to which routine tasks are delegated (leveling reading texts, providing first-draft feedback, generating vocabulary lists), the teacher can focus on the processes that require genuine expertise: pedagogical planning and individual guidance (Guilbault et al., 2025; Siegle, 2024).

The presence of AI, however, makes the teacher's role of oversight and design in the classroom more critical than ever. Given the risks that AI-generated information may contain "hallucinations" (presenting inaccurate information as if it were true) and algorithmic bias, the teacher's most important task is to serve as the epistemic and pedagogical monitor of AI outputs (Guilbault et al., 2025; Herold, 2022). It is still the teacher who designs the discussions, lines of inquiry, and critical-analysis processes to be carried out before and after students interact with virtual assistants (Siegle, 2025). By teaching students not only to converse with AI but also to catch inconsistencies in AI outputs, to compare the responses of different platforms, and to carry out "fact-checking," the teacher can develop critical thinking skills within the foreign language lesson (Guilbault et al., 2025; Siegle, 2023).

However central this role may be, it does not come about on its own; the value of AI in the classroom depends largely on how familiar the teacher is

with these tools and how they view them. Studies conducted with teachers who work with gifted students indicate that these teachers are generally open and positive toward AI (Yüreğilli Göksu & Göksu, 2024). Yet whether a teacher sees AI as a threat or as a partner is closely tied to the educational outlook they hold: a student-centered outlook that values development and individual difference tends to view AI as an opportunity, whereas a traditional outlook that prioritizes the fixed transmission of knowledge may perceive it as a threat to the teacher's own authority (İmamoğlu Akman & Akman, 2025). For this reason, the entry of AI into the classroom is not merely a matter of technical skill but also a matter of pedagogical stance; alongside tool use, teacher-preparation and professional-development programs also need to address the pedagogical outlook through which these tools become meaningful.

5.2. Differentiated AI-Mediated Task Design, Assessment, and Feedback

The teacher's role as designer and mediator finds its most concrete expression in task design. For the gifted student, differentiation is achieved not by adding a single "harder question" but through tasks reconfigured according to the student's readiness, pace, and interests, which requires a deliberate adaptation of instructional content (García-López et al., 2025). An established procedure for this adaptation is curriculum compacting: the teacher first defines the target learning outcomes, then determines which students have already mastered them, and fills the time thus freed with more advanced work (Reis & Renzulli, 2005). AI's contribution to this procedure lies in its capacity to ease the two most laborious steps—determining mastery and generating the enrichment tasks to replace it. Similarly, technology-based systems that profile students' strengths and interests and match resources accordingly hold the promise of making differentiation scalable (Renzulli & Reis, 2007).

The content of the task, in turn, should be built on the principle of depth. Prodanovska-Poposka and Todorova (2020) recommend deepening thinking through higher-order question stems and provocative "what if?" questions, and, for advanced readers, offering texts above grade level that contain demanding figures such as allegory and metaphor. In such tasks, the interaction with AI itself becomes an object of design; the teacher can treat

prompt engineering as a teachable skill and, through models such as CRAFT, model high-quality questioning for the student (Siegle, 2025). The autonomy dimension of the task should not be neglected either; environments that afford students opportunities for choice strengthen self-regulated learning (Housand & Reis, 2008). That this design can be both feasible and problematic in a real classroom has been shown by a guided-inquiry study conducted with gifted students through an AI chatbot; the study brings out both the concrete affordances and the implementation challenges together (Chang et al., 2023). An illustrative design that combines these elements can be envisioned: the teacher compacts a unit and devotes the time gained to a research task in which the student analyzes an authentic text with a generative tool and iteratively reconstructs their own prompts, while the teacher positions themselves as a mediator who monitors the quality of the questions and the student's reasoning throughout the process.

Assessment and feedback are the most delicate part of the design. AI's ability to provide immediate feedback (García-López et al., 2025) frees formative feedback and practice from depending on the teacher reaching each student one by one. Delegating the whole of assessment to the tool, however, is risky. There can be tension between teacher assessment and AI assessment (García-López et al., 2025); moreover, the human judgments used to train AI scoring systems have been found to show greater disagreement precisely on the highly original responses of gifted and bilingual students (Dumas et al., 2023). The implication for design is clear: while AI is used for formative, low-stakes feedback and practice, human judgment should be preserved for the final assessment of original and creative products.

6. Challenges, Risks, and Ethical Considerations

However notable the opportunities for personalization, acceleration, and enrichment that AI technologies offer in gifted students' foreign language education, the integration of these tools carries both affordances and risks, bringing with it pedagogical, ethical, and technical hazards (García-López et al., 2025; Guilbault et al., 2025). Gifted students' high cognitive capacity, deep intellectual curiosity, and sensitive affective makeup place them in a position that is at once more advantageous and more vulnerable to the potential adverse effects of AI (Siegle, 2023).

6.1. Pedagogical Risks: Over-Reliance, the Creativity Paradox, and Cognitive Passivity

One of the greatest pedagogical threats of AI in foreign language classrooms is the over-reliance it may foster in students and the cognitive passivity that follows from it; the ability of these tools to produce fast, fluent responses to any writing or grammar task within seconds can create a comfort zone for students (Siegle, 2023). Three manifestations of this are particularly important. In the first, students may delegate to AI the processes that require deep cognitive effort—word choice, constructing syntactic structures, and organizing ideas in a logical sequence—and thereby blunt their own research, analysis, synthesis, and problem-solving skills (Siegle, 2023). The second is the creativity paradox: although AI can spark original production, when used improperly it may steer the student toward consuming clichéd patterns distilled from average data rather than developing their own ideas, metaphors, and style; this risk is especially limiting for gifted students with high creative-productive capacity (Acar, 2023; Siegle, 2023). The third is that the artificiality of interaction with chatbots—which still lack human emotions and social cues—can be quickly noticed by students of high linguistic and cognitive level, and this can lead to disappointment and a loss of motivation in them (Chang et al., 2023; Gallacher et al., 2018; Yin & Satar, 2020).

6.2. Ethical Boundaries: Algorithmic Bias and Academic Integrity

The ethical use of AI in education is one of the prominent topics of debate in the international literature; because these models are human-made and are trained on large data sets from the internet, they may structurally harbor certain ethical flaws (Herold, 2022; Siegle, 2024). On the one hand, these systems may carry the racial, gender, socioeconomic, and cultural biases present in the data on which they are trained and reflect them in their outputs; in the foreign language classroom, this can appear as the prominence of a uniform language culture that disregards cultural diversity (García-López et al., 2025; Herold, 2022; Siegle, 2024). On the other hand, one of GenAI's best-known technical weaknesses—hallucinations, that is, the

presentation of false or fabricated information in a convincing manner—can mislead gifted students in their knowledge-based inquiries; for this reason, teaching students that AI outputs are not absolute truths and must be evaluated critically is an ethical necessity (Guilbault et al., 2025; Herold, 2022; Siegle, 2024, 2025).

A second ethical concern is academic integrity. The ability of generative models to produce seemingly original texts within seconds strains the traditional understanding of academic integrity; gifted students, when they do not find in-school assignments sufficiently challenging or when they are under time pressure, may be inclined to present generated text as their own work (Guilbault et al., 2025; Siegle, 2023). The functional solution here is not to ban AI altogether—which is often ineffective—but to teach students clearly the ethical limits of the tool, as a partner for collaboration and brainstorming rather than an instrument of plagiarism (Guilbault et al., 2025; Siegle, 2023, 2025).

6.3. Technical Barriers and Data Security: COPPA and Children's Rights

The most important factor limiting AI's use in the classroom is students' data privacy and digital security. Because GenAI platforms can record and store on their servers the data users enter (prompts, voice recordings, texts, personal information), there arises a risk that children's digital footprints may be collected without their consent (Guilbault et al., 2025). Indeed, the U.S. federal law COPPA (Children's Online Privacy Protection Act) and similar regulations (e.g., the EU's GDPR) restrict the collection of the personal data of children under 13 without parental consent; most widely used tools likewise prohibit direct account creation for those under 13 and require parental permission for ages 13 to 18 (Guilbault et al., 2025). For this reason, the literature recommends, especially with gifted students under 13, a "teacher-facing" use: rather than students using the tool directly, the teacher delivers to them the personalized content (reading texts, enriched assignments) produced through an interface under the teacher's own control, and students' names, school information, or performance data are not entered into the system (Guilbault et al., 2025).

6.4. Inequity of Access

For gifted students, the most insidious risk may well be inequity of access. None of the affordances described above come into play unless a student is identified as gifted and placed in an appropriate program. Yet the equity dimension of identification remains unresolved; the overlooking of gifted students in disadvantaged contexts means that they never gain access to these tools at all (Renzulli, 2005a). The emphasis that differentiation must be addressed separately for culturally diverse and talented students reflects the same concern (Renzulli & Reis, 2006). When technological inequity of access is added to this, the achievement and developmental gap—the excellence gap—between high-socioeconomic-status gifted students who have access to the advanced enriching resources AI offers and their disadvantaged peers who are deprived of these technologies risks widening still further (Alaca, 2018; Hodges & Mohan, 2019; Siegle, 2024). Ethically, when AI integration is planned, schools and policymakers bear the responsibility of ensuring that all gifted students can access these advanced personalized tools on equal terms (David et al., 2023; Siegle, 2024).

7. Conclusion

This chapter has sought to bring together, within a single framework, two areas that are most often treated separately: the education of gifted students and the use of AI in foreign language instruction. In doing so, AI has been positioned not as an end in itself or as a novelty, but as a contemporary response to the needs of gifted language learners—namely, pace, depth, autonomy, and interest-based production. Yet none of these affordances comes about on its own. The same tool can, in the hands of a good teacher, carry the gifted student forward; used carelessly, it can merely distribute tasks prepared for the average more quickly, widen existing inequities, and even blunt the very thinking skills the gifted student needs to develop.

For this reason, the claim in the chapter's title is not a celebration of technology but a design principle. What makes AI valuable for the gifted student is not the tool itself, but its being deliberately directed toward that student's needs for pace, depth, autonomy, and interest. Used well, AI becomes not a substitute that replaces the teacher but a thinking partner that works under the teacher's guidance.

Finally, the nature of the proposals offered here should be stated clearly. These are not practices proven effective in studies with gifted students, but possibilities constructed on the basis of established theories in gifted education. Research in this area is still new and limited. Accordingly, the aims of future studies might be to translate these possibilities into designs that can be tested in the classroom and to investigate their effects directly, particularly with gifted students.

References

Acar, S. (2023). Does the task structure impact the fluency confound in divergent thinking? An investigation with TTCT-Figural. *Creativity Research Journal*, *35*(1), 1–14.

Alaca, Z. (2018). Excellence gaps in education: Expanding opportunities for talented students [Review of the book *Excellence gaps in education: Expanding opportunities for talented students, by J. A. Plucker & S. J. Peters]. Roeper Review,* 40(1), 67–69

Bain, S. K., McCallum, R. S., Bell, S. M., Cochran, J. L., & Sawyer, S. C. (2010). Foreign language learning aptitudes, attitudes, attributions, and achievement of postsecondary students identified as gifted. *Journal of Advanced Academics, 22*(1), 130–156.

Chang, J., Park, J., & Park, J. (2023). Using an artificial intelligence chatbot in scientific inquiry: Focusing on a guided-inquiry activity using Inquirybot. *Asia-Pacific Science Education, 9*(1), 44–74.

David, A., Kiose, V., Maikou, A., Tzelepi, E., & Stathopoulou, A. (2023). The impact of ICTs (robotics, VR, AI, games) on gifted students' education. *Eximia*, *8*(1), 31–50.

Dumas, D., Acar, S., Berthiaume, K., Organisciak, P., Eby, D., Grajzel, K., Vlaamster, T., Newman, M., & Carrera, M. (2023). What makes children's responses to creativity assessments difficult to judge reliably? *The Journal of Creative Behavior*, *57*(3), 419–438.

Gallacher, A., Thompson, A., Howarth, M., Taalas, P., Jalkanen, J., Bradley, L., & Thouësny, S. (2018). "My robot is an idiot!" – students' perceptions of AI in the L2 classroom. In P. Taalas, J. Jalkanen, L. Bradley, & S. Thouësny (Eds.), *Future-Proof CALL: Language learning as exploration and encounters–Short papers from EUROCALL 2018* (pp. 70–78). Research-publishing.net.

García-López, C., Tabuenca-Cuevas, M., & Navarro-Soria, I. (2025). A systematic review of the use of AI in EFL and EL classrooms for gifted students. *Trends in Higher Education, 4 (3)*, Article 33.

Guilbault, K. M., Wang, Y., & McCormick, K. M. (2025). Using ChatGPT in the secondary gifted classroom for personalized learning and mentoring. *Gifted Child Today, 48*(2), 93–103.

Herold, B. (2022, April 12). Why schools need to talk about racial bias in AI-powered technologies. *Education Week*.

Hodges, J., & Mohan, S. (2019). Machine learning in gifted education: A demonstration using neural networks. *Gifted Child Quarterly*, *63*(4), 243–252.

Housand, A. M., & Reis, S. M. (2008). Self-regulated learning in reading: Gifted pedagogy and instructional settings. *Journal of Advanced Academics, 20*(1), 108–136.

İmamoğlu Akman, G., & Akman, Y. (2025). Educational beliefs and artificial intelligence: A study on pre-service teachers [Eğitim inançları ve yapay zekâ: Öğretmen adayları üzerine bir araştırma]. *Batı Anadolu Eğitim Bilimleri Dergisi*, *16*(1), 911–928.

Prodanovska-Poposka, V., & Todorova, M. (2020). Foreign language teaching for gifted learners: Overview of listening, reading, speaking and writing. In *Gifted and talented creators of the progress: Conference proceedings* (pp. 552–557). St. Kliment Ohridski University–Bitola, Faculty of Education.

Reis, S. M., & Renzulli, J. S. (2005). *Curriculum compacting: A systematic procedure for modifying the curriculum for above average ability students*. University of Connecticut, The National Research Center on the Gifted and Talented.

Reis, S. M., Renzulli, J. S., & Burns, D. E. (2016). *Curriculum compacting: A guide to differentiating curriculum and instruction through enrichment and acceleration* (2nd ed.). Prufrock Press.

Renzulli, J. S. (2011). What makes giftedness? Reexamining a definition. *Phi Delta Kappan, 92*(8), 81–88. (Original work published 1978)

Renzulli, J. S. (2005a). *Equity, excellence, and economy in a system for identifying students in gifted education programs: A guidebook* (Research Monograph RM05208). University of Connecticut, The National Research Center on the Gifted and Talented.

Renzulli, J. S. (2005b). The three-ring conception of giftedness: A developmental model for promoting creative productivity. In R. J. Sternberg & J.

E. Davidson (Eds.), *Conceptions of giftedness* (2nd ed., pp. 246–279). Cambridge University Press.

Renzulli, J. S. (2012). Reexamining the role of gifted education and talent development for the 21st century: A four-part theoretical approach. *Gifted Child Quarterly, 56*(3), 150–159.

Renzulli, J. S. (2016). The three-ring conception of giftedness: A developmental model for promoting creative productivity. In S. M. Reis (Ed.), *Reflections on gifted education* (pp. 55–86). Prufrock Press.

Renzulli, J. S., & Reis, S. M. (1997). *The Schoolwide Enrichment Model: A how-to guide for educational excellence* (2nd ed.). Creative Learning Press.

Renzulli, J. S., & Reis, S. M. (2006). Curriculum compacting: A research-based differentiation strategy for culturally diverse talented students. In B. Wallace & G. Eriksson (Eds.), *Diversity in gifted education* (pp. 73–85). Routledge.

Renzulli, J. S., & Reis, S. M. (2007). A technology based program that matches enrichment resources with student strengths. *International Journal of Emerging Technologies in Learning, 2*(3), 1–12.

Renzulli, J. S., & Reis, S. M. (2014). *The Schoolwide Enrichment Model: A how-to guide for talent development* (3rd ed.). Prufrock Press.

Serafeim, M., & Alexiou, T. (2020). Investigating foreign language aptitude in intellectually gifted students. In W. Zoghbor & T. Alexiou (Eds.), *Advancing English language education* (pp. 125–142). Zayed University Press.

Siegle, D. (2023). A role for ChatGPT and AI in gifted education. *Gifted Child Today, 46*(3), 211–219.

Siegle, D. (2024). Using artificial intelligence (AI) technology to support the three legs of talent development. *Gifted Child Today, 47*(4), 221– 227.

Siegle, D. (2025). Using AI prompt engineering to improve gifted students' questioning. *Gifted Child Today, 48*(1), 68–72.

Sun, H., Li, S., Kim, D., & Tang, S. (2026). The role of language aptitude in child second language learning: A meta-analysis. *Educational Psychology Review, 38 (1)*, Article 37.

Tomlinson, C. A. (2014). Differentiated instruction. In J. A. Plucker & C. M. Callahan (Eds.), *Critical issues and practices in gifted education: What the research says* (2nd ed., pp. 197–210). Prufrock Press.

Wei, L. (2023). Artificial intelligence in language instruction: Impact on English learning achievement, L2 motivation, and self-regulated learning. *Frontiers in Psychology, 14*, Article 1261955.

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry, 17*(2), 89–100.

Yin, Q., & Satar, M. (2020). English as a foreign language learner interactions with chatbots: Negotiation for meaning. *International Online Journal of Education and Teaching*, *7*(2), 390–410.

Yüreğilli Göksu, D., & Göksu, S. (2024). Determination of artificial intelligence literacy and attitudes towards artificial intelligence of teachers working with gifted students and examining them according to some variables. *Instructional Technology and Lifelong Learning*, *5*(2), 278–299.

Zhang, Q., Nie, H., Fan, J., & Liu, H. (2025). Exploring the dynamics of artificial intelligence literacy on English as a foreign language learners' willingness to communicate: The critical mediating roles of artificial intelligence learning self-efficacy and classroom anxiety. *Behavioral Sciences*, *15*(4), Article 523. https://doi.org/10.3390/bs15040523

From Algorithm to Voice: The Pedagogical Potential and Limitations of AI-Generated Spoken Content

Doğa Selin Geçer[1] & *Yusuf Topaloğlu*[2]

1. Introduction

Artificial intelligence (AI) has become increasingly integrated into many aspects of everyday life. The foundations of the field were laid in 1950, when Alan Turing proposed the "Imitation Game" to examine whether machines could imitate human behavior; the term "artificial intelligence" itself was first coined by John McCarthy at the Dartmouth Conference in 1956. Today, artificial intelligence has evolved, through its capacity for deep learning and its ability to recognize complex patterns, into a technology capable of transforming numerous sectors by means of artificial neural networks. Defined as a machine's capacity to imitate cognitive processes specific to human intelligence (such as learning, reasoning, problem-solving, and decision-making), artificial intelligence has ceased to be merely a technology that processes data; it has become a system capable of generating content, interacting with users, and producing creative outputs across diverse domains. In particular, the development of generative artificial intelligence (GenAI) models has made it possible to automatically produce a wide range of content, including text, images, video, and audio.

One of the most notable application areas of generative AI is voice-generation technology. Technologies such as text-to-speech (TTS), speech synthesis, and

[1] Lecturer, National Defence University, ORCID: 0009-0004-2918-7961, *dogaselingecer@gmail.com*, Türkiye

[2] Assoc. Prof. Dr., Istanbul University Cerrahpaşa, Hasan Âli Yücel Faculty of Education, Department of Foreign Language Education, Division of French Language Education, ORCID: 0000-0003-2832-6902, *yusuf.topaloglu@iuc.edu.tr,* Türkiye

voice cloning can convert written input into natural, human-like voices, enabling the production of spoken content with different languages, accents, and speech characteristics. The integration of deep-learning-based models and large language models (LLMs) with voice technologies has considerably enhanced the naturalness and interactive capabilities of AI-generated voice systems. AI-supported voice systems have thus evolved from a mere technical novelty into functional tools employed across numerous fields, including education, media, healthcare, and accessibility.

Education is one of the fields in which this transformation is most intensely felt. With the spread of digital learning environments, spoken content has become an important component of the learning process, and AI-supported voice technologies have reshaped the ways in which such content is produced and delivered. Audio materials that can be adapted to learners' needs render educational processes more flexible, accessible, and personalizable. In particular, the capacity of these technologies to reduce constraints of time and place, allow learners to work at their own pace, and provide content suited to diverse learning profiles increases their importance in education.

When considered in the context of foreign language teaching, AI-generated spoken content offers a wide range of applications, from the development of listening and pronunciation skills to the creation of original dialogues, the introduction of different accents, and the design of personalized learning experiences. In addition, it functions as an important accessibility tool for students with visual impairments, individuals with reading difficulties, or learners in distance-education settings. Owing to these features, voice-generation technologies stand out as innovative tools that support learner autonomy and strengthen equity of opportunity in education.

Nevertheless, the use of AI-generated spoken content in educational settings should be evaluated not only in terms of technical competence but also regarding pedagogical appropriateness, ethical dimensions, and contribution to the learning process. This chapter first addresses the theoretical foundations of AI-based voice-generation technologies and then examines their areas of application in education and their pedagogical potential. Finally, it discusses the limitations of AI-based voice systems, the ethical debates surrounding them, and considers future directions for development by proposing an analytical framework for their evaluation in educational settings.

2. Theoretical Foundations of AI-Based Voice-Generation Technologies

2.1 From Artificial Intelligence to Generative Artificial Intelligence

In recent years, advances in the field of artificial intelligence have driven progress across numerous domains, including healthcare, media, and education. Because AI has gained particular momentum in tasks that traditionally required human creativity (such as text, image, and voice generation), new content can now be produced rapidly and efficiently. These systems are generally capable of converting written text into natural, fluent speech. Artificial intelligence can autonomously perform a requested task, at a requested time, place, and pace, based on the commands given to it, thereby enabling everyday tasks to be completed in considerably less time and effort. However, while simple AI tools continue to perform their existing functions today, they remain unable to go beyond following commands, and this limitation has paved the way for the need for, and the development of, generative AI tools. Then, what are these simple AI tools? They can first be classified into three types: narrow AI, general AI, and superintelligent AI.

2.1.1. Narrow AI

Narrow AI tools are those that serve a single function and are used only within limited domains. The first examples that come to mind under this category are virtual assistants, the best known of which are Apple's Siri and Amazon's Alexa. Search engines such as Google and Yandex are likewise examples of AI within this category. Applications used mainly for entertainment, such as Netflix and Spotify, also draw on this type of AI in the form of recommendation algorithms. Language-translation tools such as DeepL and Google Translate, which are widely used especially in foreign language learning, belong to this group as well. In addition, email filters that sort incoming messages into spam folders, chatbots that assist consumers experiencing problems with purchases, and systems equipped with image and facial recognition capabilities all constitute further examples of narrow AI applications.

2.1.2. General AI

It refers to a hypothetical form of artificial intelligence capable of performing a wide range of cognitive tasks at a level comparable to human intelligence. Current systems such as ChatGPT, Midjourney, and Notion AI demonstrate advanced generative capabilities, but they are not generally considered examples of artificial general intelligence.

Superintelligent AI (Artificial Superintelligence – ASI)

Artificial superintelligence (ASI) refers to a hypothetical form of AI that would surpass human intelligence across nearly all cognitive domains. No system currently recognized as ASI exists. Advanced systems such as AlphaGo and AlphaFold demonstrate exceptional performance in specific domains, but they remain examples of narrow AI. AlphaGo was developed to play the game of Go, while AlphaFold was designed to predict protein structures. Although research continues to expand the capabilities of AI systems in areas such as coding, data analysis, and reasoning, these developments should not be regarded as evidence that artificial superintelligence has already been achieved.

2.1.3. Generative Artificial Intelligence (GenAI)

When the different uses and capabilities of artificial intelligence are examined, it becomes evident that a substantial portion of these systems is designed to perform a specific task, analyze existing data, classify information, make predictions, or carry out a particular operation in response to a user-issued command. While this tendency is more pronounced in narrow AI applications, the concepts of general AI and superintelligent AI are approached in terms of the imitation, or transcendence, of human-like cognitive abilities at different levels instead. Nevertheless, developments in the field of artificial intelligence in recent years have moved beyond systems that merely process existing information or perform a designated task, giving rise to the proliferation of AI models capable of generating new content. At this juncture, generative artificial intelligence (GenAI) emerges as a distinct conceptual domain, referring to AI's capacity to produce new and meaningful content based on patterns learned from existing data (Feuerriegel et al., 2024).

The defining characteristic of generative AI is that it is not confined to merely analyzing the data provided to it or arriving at a previously defined outcome; rather, it is capable of generating a new output based on the patterns it has learned. In other words, whereas the core function of traditional AI applications is typically to "derive a conclusion from existing data," in generative AI this process instead takes the form of "producing new content on the basis of existing data." During training, generative models learn linguistic, visual, auditory, or other patterns embedded within vast quantities of data and, on this basis, are able to generate outputs that did not previously exist in identical form, guided by user-provided input. Generative AI is therefore regarded not merely as a tool for information retrieval or automation, but as a technology that participates directly in the process of content creation (Feuerriegel et al., 2024).

Another point that distinguishes generative AI from other AI applications is the nature of its output. A recommendation system, for instance, may suggest particular content by analyzing a user's past behavior; a translation system may render a given text into another language; or an image-recognition system may classify the objects present in an image. The primary purpose of such systems is to process existing inputs and produce a resulting output. In generative AI systems, however, a text prompt, an image, an audio sample, or another form of data input provided by the user can serve as the starting point for the creation of new content. The system thus ceases to function merely as an entity that "recognizes," "classifies," or "recommends," and instead becomes, to a certain extent, a generative structure. This generative capacity can be applied to diverse content types, including text, images, music, audio, video, and software code (Feuerriegel et al., 2024; Autio et al., 2024).

The development of large language models (LLMs) and other foundation models has played a significant role in the increasing visibility of generative AI in recent years. Trained on vast quantities of data, these models can process natural-language commands provided by users to generate contextually appropriate text, answer questions, summarize texts, translate between different languages, and produce various forms of written content. Likewise, generative models are not limited to textual content alone. Visual generation models can produce images from given descriptions, while

audio and music generation systems can produce new spoken or musical outputs from textual or auditory inputs. Generative AI, therefore, does not refer to a single application or a single data type, but rather encompasses a broad family of technologies capable of generating new content across different data modalities (Feuerriegel et al., 2024).

At this point, it is also important to distinguish generative AI from the concept of artificial general intelligence (AGI). Whereas generative AI refers to an approach centered on the capacity to generate new content, artificial general intelligence denotes a hypothetical level of AI capable of performing a broad range of human cognitive tasks, unrestricted to any specific task or domain. Accordingly, classifying present-day systems such as ChatGPT, image-generation systems, or voice-generation tools under the umbrella of generative AI does not imply that they have attained human-level general intelligence. Indeed, although generative AI systems are capable of producing highly sophisticated content, the accuracy, reliability, and contextual appropriateness of their outputs still require human evaluation. UNESCO likewise emphasizes the importance of preserving a human-centered approach in the use of generative AI in education and research, and of critically evaluating technological output (UNESCO, 2023).

The emergence of generative AI should not, therefore, be understood only as artificial intelligence becoming "more intelligent." The more significant transformation lies in AI's expansion from a tool that processes data into one that participates in the content-generation process. This does not mean, however, that increased generative capacity renders these systems limitless or error-free. Generative AI models may produce content that is inaccurate, incomplete, biased, or contextually inappropriate, and they may also raise a range of concerns regarding data privacy, copyright, transparency, and reliability. For this reason, the evaluation of GenAI technologies should attend not only to the technical quality of the resulting content, but also to the data and models through which that content was produced, the purpose for which it is used, and the manner in which the resulting output is assessed by human evaluators (Autio et al., 2024; UNESCO, 2023).

This generative capacity of AI carries particular significance for voice technologies specifically. As in text generation, AI systems in the domain of speech are not limited to analyzing an existing audio recording; they can

convert textual input into speech, model different speaker characteristics, and generate new spoken content with specified properties. Text-to-speech (TTS), speech synthesis, and voice cloning can thus be regarded as manifestations of generative AI in the domain of voice. From this standpoint, understanding the general characteristics of generative AI provides a fundamental framework for assessing how AI-based voice-generation technologies function and what pedagogical opportunities they offer within educational settings.

2.2. Text-to-Speech (TTS) Systems

One of the most widespread applications of AI-supported voice technologies in education is text-to-speech (TTS) systems, which automatically convert written text into speech. In its broadest sense, TTS refers to the process of generating speech from written input and is regarded as one of the principal means of providing learners with spoken linguistic input within computer-assisted language-learning environments (Handley, 2009). With the development of these systems, it has become possible not solely to read written content, but to have it voiced at different speeds, with different vocal characteristics, and in a repeatable manner.

The mechanical, monotonous voices produced by traditional TTS systems were markedly distinguishable from human speech, particularly with respect to natural speech rhythm, stress, and intonation. However, the development of deep-learning-based speech-synthesis models has brought about a significant transformation in this regard. Tacotron 2, for example, developed a neural architecture that generates speech directly from written input and has been shown to produce voices whose naturalness closely approaches that of professional human recordings (Shen et al., 2018). Similarly, FastSpeech 2 was designed to directly model speech characteristics such as pitch, energy, and duration, with the aim of producing faster and higher-quality TTS output (Ren et al., 2021).

This technological progress has also rendered TTS systems more functional for foreign language education. TTS applications can be particularly beneficial in helping students notice sounds in the target language, improve their pronunciation, and gain greater exposure to listening input. Indeed, in a study conducted by Liakin, Cardoso, and Liakina (2017) with learners of

French, mobile TTS technology was found to support pronunciation practice aimed at developing liaison production in French. In this study, students engaged in listening, categorizing, and repetition activities through TTS, and the researchers concluded that TTS could be used as a complementary tool in second-language pronunciation instruction.

The function of TTS in education is not limited to language teaching alone. The ability to convert written materials into spoken form facilitates access to educational content, particularly for students with visual impairments or reading difficulties. Similarly, Bone and Bouck (2017) emphasize that TTS applications can support independent access to written content for students with reading difficulties. For this reason, TTS can be regarded not only as a text-to-speech conversion technology, but also as an assistive technology that contributes to the creation of accessible and inclusive learning environments.

2.3. Speech Synthesis

Closely related to TTS yet representing a broader technological domain, speech synthesis refers to the artificial process of generating a voice that resembles human speech. Whereas TTS specifically denotes the conversion of a given written input into speech, speech synthesis encompasses the broader linguistic, acoustic, and signal-processing processes underlying this conversion. In particular, the use of deep neural networks and sequence-to-sequence learning architectures has made it possible for speech synthesis to produce more natural and intelligible voices (Shen et al., 2018).

One significant example of this transformation, Tacotron 2, introduced an architecture that first converts textual input into mel-spectrograms and then transforms this representation into a speech waveform via a WaveNet-based vocoder. The results obtained in this study demonstrated that synthetic speech could attain a level of naturalness closely approaching that of professional human speech (Shen et al., 2018). Later models placed increasing emphasis on modeling not only the lexical content of speech, but also variables such as pitch, energy, duration, and other prosodic features. In this respect, FastSpeech 2 serves as an important example by modeling variations in speech more directly, thereby improving both synthesis speed and voice quality (Ren et al., 2021).

The range of applications for speech synthesis has likewise expanded alongside these technological advances. Audiobooks, virtual assistants, automated narration systems, and human–machine interactions are among the principal application domains of this technology. In educational contexts, speech synthesis can be used for a variety of purposes, including the narration of digital materials, the development of interactive learning environments, and enabling students to participate in the learning process through spoken input. In this respect, speech synthesis contributes to the creation of voice-based and interactive learning experiences, moving educational content beyond the confines of written and visual materials alone.

Nevertheless, the fact that synthetic speech has come to approximate human speech does not mean that it is fully equivalent to it. In particular, the accurate modeling of elements such as emotional expression, contextual intonation, and pragmatic meaning remains an important area of ongoing research. Consequently, although technical progress in speech synthesis has enabled increasingly natural voice production, its educational use requires evaluating not only the acoustic quality of the voice, but also its contextual appropriateness and pedagogical purpose.

2.4. Voice Cloning

Representing a further stage in the development of voice technologies, voice cloning refers to the process of analyzing a specific individual's vocal characteristics in order to construct an artificial voice model that resembles that person. Unlike conventional TTS systems, this technology enables not simply the generation of a natural synthetic voice, but the creation of entirely new speech by modeling the vocal characteristics of a specific speaker. The ability to model pitch, tone, speech rhythm, and other acoustic properties of a voice has paved the way for personalized voice generation.

From an educational perspective, voice cloning offers notable opportunities in that it allows teachers to use their own voices in digital instructional materials. For example, modeling a teacher's vocal characteristics may make it possible to produce new course content or updated materials in a similar voice without the need for re-recording. Furthermore, hearing a familiar teacher's voice across different learning materials may be regarded as a factor that contributes to the personalization of educational content.

However, the opportunities offered by voice-cloning technology also entail significant ethical and security concerns. The capacity of AI-based systems to imitate human voices with considerable realism creates the potential for a voice to be used without its owner's knowledge or consent, and for fabricated audio content to be produced. In recent years, in particular, artificially generated voice content referred to as audio deepfakes has been associated with risks such as fraud, identity impersonation, disinformation, and security breaches (Li et al., 2025). Recent research indicates that TTS and voice-conversion technologies can be used to produce realistic fake voices, making the detection of audio-based deepfakes an increasingly important concern.

For this reason, the use of voice cloning in education must consider principles such as explicit consent, the protection of voice data, the intended purpose of use, the safeguarding of digital identity, and the clear disclosure that content has been artificially generated. In other words, the pedagogical value of voice-cloning technology should be assessed not only in terms of its technical capacity, but also with respect to its ethical and security dimensions.

2.5. Large Language Models (LLMs) and Voice Integration

The transformation in voice technologies has not been confined to advances in TTS and speech-synthesis models alone; it has reached a new stage through the integration of large language models (LLMs) with voice technologies. Large language models are AI systems trained on vast datasets that are capable of analyzing natural language, interpreting meaning, and generating new textual content. The integration of these models with voice technologies represents a shift away from AI-only narration of a pre-prepared text, toward a system capable of understanding a user's input, generating an appropriate response, and delivering that response in spoken form.

Recent research classifies the integration of LLMs with speech technologies broadly into text-based, implicit/latent-representation-based, and direct audio-token-based approaches (Yang et al., 2025). This development enables AI systems to move beyond a text-centered orientation, allowing them to directly process more natural forms of human communication, such as speech and voice.

This integration holds particular significance for education. Whereas the fundamental process in a conventional TTS system follows a "text →

speech" pathway, in systems that combine LLMs with voice technologies the process becomes considerably more interactive: "user input → interpretation → response generation → spoken output." Learners are thereby able to engage in reciprocal communication with AI, rather than merely listening to a pre-prepared text. For instance, a foreign language learner may converse with an AI system on a given topic, receive a response appropriate to their proficiency level, and continue the same conversation at different speeds or with different content.

This capacity holds distinct promise for foreign language teaching. Enabling learners to engage in spoken interaction in the target language may allow listening and oral-production skills to be exercised together. Indeed, recent research demonstrates that the use of TTS-based systems in foreign language listening instruction can support students' listening skills (Xiong, 2026). Nevertheless, the naturalness of AI-generated speech alone is not sufficient; factors such as proficiency level, linguistic accuracy, pronunciation, context, pragmatic appropriateness, and cultural content must also be taken into account.

In sum, when TTS systems, speech synthesis, voice cloning, and LLM–voice integration are considered together, it becomes evident that AI-based voice generation has evolved from a simple text-reading function into a personalizable and interactive communication technology. This transformation offers significant opportunities, particularly with respect to foreign language teaching, accessibility, and personalized learning. At the same time, in addition to vocal naturalness and technical quality, factors such as data security, ethical use, voice cloning, deepfakes, cultural representation, and pedagogical appropriateness must also be taken into consideration. The educational value of AI-generated spoken content should therefore be assessed not solely on the basis of how "human-like" it sounds, but rather in terms of the pedagogical purpose, the context, and the ethical framework within which it is used.

2.6. The Development of Text-to-Speech Technologies

The development of text-to-speech (TTS) technologies emerged from decades of research aimed at enabling machines to reproduce human speech. The fundamental purpose of TTS is to convert written text into a computer-

processable linguistic representation and, from that representation, to generate a voice resembling human speech. Yet the emergence of today's remarkably natural-sounding synthetic voices was not the outcome of a sudden technological breakthrough. On the contrary, TTS systems have passed through several successive stages: rule-based synthesis, formant synthesis, concatenative synthesis, statistical methods, and, ultimately, deep-learning-based models (Klatt, 1987).

2.6.1. Early TTS systems

The origins of speech synthesis extend well beyond the advent of digital computers. The first mechanical and electronic speech synthesizers were developed in the mid-twentieth century, with research on formant and articulation-based synthesizers conducted as early as the 1950s. In systems developed during this period, the formation of the human voice, the vocal tract, and acoustic characteristics were modeled mathematically. For example, systems such as PAT (Parametric Artificial Talker) and OVE (Orator Verbis Electris), developed in 1953, are early examples that produced speech through acoustic and articulatory modeling.

One of the key milestones in the history of TTS was the first fully automatic text-to-speech system, developed in Japan in 1968 by Noriko Umeda and colleagues. The system attempted to generate speech by analyzing the linguistic properties of the input text; although the resulting voice was intelligible, it was, by contemporary standards, considerably monotonous and unnatural in character. With the development of systems such as MITalk and, later, Klattalk in the 1970s, text analysis, phoneme conversion, and the modeling of the acoustic properties of speech became considerably more sophisticated (Klatt, 1987).

The TTS systems of this period relied largely on a rule-based logic of operation. The system would first parse the written text into linguistic units such as words and phonemes, and would then generate a speech signal using predefined rules governing pronunciation, duration, pitch, and stress. Although this approach made it possible for a computer to model the basic characteristics of speech, it struggled to adequately capture the variability and context-dependent features of natural human speech. As a result,

synthetic voices of this early period were largely mechanical, monotonous, and readily distinguishable from human speech (Klatt, 1987).

In the 1980s and 1990s, a different approach came to the fore in TTS technologies. In addition to rule-based systems, concatenative synthesis (based on combining segments taken from natural speech recordings) and, subsequently, statistical methods were developed. Although these methods enabled the production of more natural-sounding output, they required large quantities of recorded speech data and detailed linguistic-acoustic models. As a result, the focus of TTS research shifted gradually from only "producing speech" toward producing speech that was increasingly natural, intelligible, and fluent.

2.6.2. The transition to neural TTS systems

One of the most significant transformations in TTS technology occurred with the introduction of machine learning, and of deep neural networks, into speech synthesis. Deep-learning-based models can learn the complex acoustic characteristics of human speech from large datasets, substantially reducing the need for manually predefined rules. This transformation brought about important advances not only in the naturalness of synthetic voices, but also in the modeling of features such as speech rhythm, stress, pitch, and intonation.

Among the notable examples of this period are Tacotron and its enhanced successor, Tacotron 2. The Tacotron approach introduced an end-to-end neural architecture that converts textual input directly into an acoustic representation of speech. Tacotron 2, in turn, generated mel-spectrograms from text via a neural network and converted this representation into a speech signal through a WaveNet-based vocoder. The study by Shen and colleagues (2018) demonstrated that this approach could produce voices closely approximating human speech, marking one of the significant milestones in the history of neural TTS.

The development of neural TTS did not stop there. Subsequent models sought not only to improve voice quality but also to increase synthesis speed and provide greater control over the prosodic features of speech. FastSpeech 2 represents an important example of this progress; by directly modeling variables such as pitch, energy, and speech duration, it aimed to

produce faster and higher-quality speech. TTS systems thus evolved from systems that just concatenated pre-recorded speech segments into neural systems capable of learning diverse speech characteristics and generating novel, original vocal output.

The principal outcome of this transformation has been a gradual narrowing of the perceptual gap between artificial and human voices. Whereas the presence of a distinctly "machine voice" was highly noticeable in early TTS systems, in neural TTS models the fluency, stress, and intonation of speech have become considerably more natural. The fundamental shift in the development of TTS technology can thus be characterized as a transition from rule-based speech generation to speech generation that learns from data.

2.6.3. Current TTS tools

Today, TTS technology is no longer confined to experimental use within research laboratories; it has become part of commercial, cloud-based applications directly accessible to users. The advancement of generative AI models has played a significant role in this transformation. Current systems not only convert written text into speech but also offer features such as diverse vocal characters, speaking styles, multiple languages, and, in certain cases, real-time interaction.

One current example in this domain is the ElevenLabs platform, which offers multilingual text-to-speech models alongside voice-design and voice-cloning capabilities. Among ElevenLabs' current models are fast models designed for real-time use and multilingual TTS models supporting numerous languages, including French and Turkish. These features illustrate how TTS technology has expanded from systems that merely narrate a standard text into generative voice platforms adaptable to diverse vocal characteristics and use scenarios.

Another significant example is OpenAI's voice technologies. OpenAI's TTS models can be used to convert textual input into natural spoken output. The current API framework includes various TTS models and voice options, with more advanced functions such as custom voice generation also available under certain usage conditions. Moreover, OpenAI's current voice technologies are not limited to conventional text-to-speech conversion alone. As of 2026, newly developed voice models have begun to

combine real-time speech processing, response generation, and translation functions within the same interactive environment.

Google Cloud Text-to-Speech is likewise an important example within the current TTS ecosystem. In addition to standard and WaveNet-based voices, the platform includes Neural2 and newer generative models. Google's current Chirp 3: HD voices aim to model the intonation characteristics of human speech in greater detail and to produce low-latency speech for real-time applications. The platform also offers distinct synthetic voices for numerous languages, including Turkish and French.

Considered together, these developments suggest that the evolution of TTS technology can be summarized in three fundamental stages: rule-based systems, statistical and concatenative synthesis systems, and neural/generative TTS systems. Whereas the primary aim in the earliest period was the intelligible narration of written text, the goal today is increasingly oriented toward natural, fluent, multilingual, personalizable, and interactive voice generation. The development of TTS technology should therefore be understood not merely as technical progress in voice quality, but as a restructuring of the relationship between text and human-like speech.

This historical and technical transformation has also directly influenced the ways in which TTS is used in education. In particular, the increasing approximation of neural TTS and generative voice systems to natural speech has made it possible not merely to narrate written educational content, but to produce spoken content that can be personalized and adapted to pedagogical purposes.

3. The Evolution of Voice Technologies in Educational Settings

The use of voice technologies in education began long before the advent of digital AI applications and evolved in parallel with broader transformations in educational technology. In foreign language teaching in particular, voice has been one of the principal tools enabling learners to be exposed to the target language and to engage with auditory input beyond written materials. In this respect, the development of voice technologies in education extends beyond changes in the devices employed. It also involves a transformation

in how spoken input is delivered, accessed, repeated, and incorporated into the learning process.

3.1. Cassettes and recording-based audio materials

One of the key stages in the spread of voice technologies in education was the incorporation of magnetic-tape and cassette technology into instructional materials. In particular, the audio cassettes that accompanied foreign language textbooks in the latter half of the twentieth century enabled students to listen to target-language speech both inside and outside the classroom. Through these materials, students gained access to auditory input that was not limited to the teacher's speech within the classroom. These recordings could also be used to conduct dialogue, vocabulary, pronunciation, and listening activities.

One of the most significant features of cassette technology from an educational standpoint was that audio content could be reused. Students could listen to a given recording multiple times, allowing repeated exposure to the same content, particularly in pronunciation or listening exercises. Nonetheless, cassette-based systems remained dependent on physical materials; modifying content, reproducing it, or accessing a specific segment directly was considerably more limited than in today's digital systems. This period can thus be regarded as the stage at which voice became established in educational settings as a recorded and reusable learning resource.

3.2. Language laboratories

Another important stage in the development of voice technologies in education was the emergence of language laboratories, particularly in connection with foreign language teaching. Language laboratories were designed as technically equipped learning environments that enabled students to listen to audio recordings and repeat particular linguistic patterns. These systems, which were widely used during the 1960s and 1970s in connection with behaviorist approaches to learning, placed listening and repetition at the core of the language-learning process (Kitao, 1995).

In language laboratories, students typically listened to a model in the target language through headphones, repeated the given utterance, and, in some systems, recorded their own voices for comparison with the model

recording. Voice technology thus began to shift from a passive listening tool toward a learning environment that also engaged students in spoken production. Nevertheless, a substantial portion of these laboratories relied on particular instructional models and repetition-based activities. As a result, unlike today's interactive digital systems, this technology largely served to convey content predetermined by the teacher to the student. Vanderplank (2010), evaluating a decade of research on language laboratories, television, and video, demonstrates that although these technologies hold potential for supporting listening skills, their effects vary depending on instructional design and mode of use.

3.3. The emergence of podcasts

With the spread of the internet and portable digital devices, the nature of voice use in education changed considerably. One notable example of this transformation is the emergence of the podcast format. By enabling spoken content to be published online and accessed by users at a time of their choosing, podcasts contributed to freeing audio materials in education from dependence on physical carriers.

In foreign language teaching, podcasts hold significant potential, particularly with respect to access to authentic and diverse linguistic input. Students can listen not only to studio recordings accompanying a textbook, but also to real-life content representing different speakers, regions, and communicative situations. This exposes students to varied speech rates, accents, voices, and communicative styles. The downloadable and portable nature of podcasts, in turn, has allowed learning to extend beyond the classroom, enabling students to structure their own study time. Indeed, the study by Abdous, Facer, and Yen (2012) shows that the integrated use of podcasts within course content contributes more strongly to academic achievement in foreign language courses than does supplemental use. Similarly, Alm (2013) demonstrates that podcasts, by facilitating access to authentic and diverse listening input, support "extensive listening" practices among foreign language learners.

During this period, the nature of spoken learning materials also changed. Whereas in the cassette era content was largely prepared in advance as part of a fixed curriculum, podcasts gave rise to a broader ecosystem of audio

content that could be continuously updated online and selected from diverse sources. Consequently, learners moved from relying on a single audio recording provided to them to exercising greater autonomy by choosing among various audio sources.

3.4. Digital learning environments

The spread of digital learning environments has once again transformed the use of voice technologies in education. Through learning management systems, online course platforms, mobile applications, and digital course materials, voice has ceased to function as a standalone recording and has instead become part of multimedia content integrated with text, visuals, video, and interactive activities.

This transformation is particularly evident in foreign language education. Students can read a text while simultaneously listening to its spoken version, replay a specific segment, and, in certain digital applications, record their own voices to practice pronunciation. Over time, the function of voice technology in education has expanded beyond its initial role of providing listening material to creating multimedia learning experiences in which students can draw on multiple skills simultaneously.

Another important feature of digital environments is that learning materials can be more readily updated and adapted to different groups of learners. Because audio recordings are no longer tied to a physical carrier, content can be accessed from a variety of devices, allowing educational materials to be used independently of time and place. Voice has therefore moved beyond its role as a supplementary classroom resource and has become a continuously accessible component of online and mobile learning experiences.

3.5. AI-supported voice systems

One of the stages currently reached in the development of voice technologies in education is the incorporation of AI-supported voice systems into learning environments. Whereas in earlier periods the voice used in educational materials was largely pre-recorded by a teacher or a professional speaker, AI-based systems make it possible to generate and modify spoken content dynamically. In particular, advances in text-to-speech (TTS)

technologies now allow written educational content to be converted rapidly into synthetic voices resembling natural speech.

At this point, the neural TTS systems discussed in the previous section become especially important. Models such as Tacotron 2 and FastSpeech 2, by enabling the acoustic characteristics of speech to be modeled through deep learning, have substantially enhanced the naturalness of synthetic voices (Shen et al., 2018; Ren et al., 2021). Teachers or materials developers can thus have written content voiced through artificial intelligence, rather than needing to produce a new audio recording for each new text.

A fundamental distinction between AI-supported systems and earlier voice technologies is that content can be generated and adapted. During the cassette or podcast eras, audio content was largely pre-recorded and fixed, whereas in AI-based systems the same content can be regenerated at different speeds, with different voices, or based on different texts. This capacity opens up new possibilities, particularly with respect to personalized learning. Examples of such possibilities include the preparation of texts suited to a student's proficiency level, the narration of texts at different speeds, or the creation of content focused on a specific pronunciation feature.

Furthermore, another feature that distinguishes AI-supported voice systems from earlier technologies is their interactive capacity. Current systems do not merely narrate pre-prepared content; through the combined use of speech recognition, natural language processing, large language models, and speech synthesis, they are capable of engaging in reciprocal spoken communication with users. Learners therefore move beyond simply listening to a fixed audio recording, becoming active participants who can converse with AI, ask questions, receive responses, and sustain interaction.

This historical transformation clearly illustrates the changes that voice technologies have undergone in education. Voice has evolved from a recorded and replayable material on cassettes, to a tool for controlled repetition and oral production in language laboratories, to a portable and accessible form of content in podcasts, to a component of multimedia in digital learning environments, and, in AI-supported systems, to a generative, adaptable, and interactive learning resource. The transformation occurring today therefore does not simply reflect technical progress in voice quality, but also signals a redefinition of the role of voice within the educational process.

This shift also illustrates why the pedagogical potential of AI-generated spoken content warrants separate consideration. Unlike previous technologies, current AI systems do not merely deliver an existing piece of audio content to a student; they can generate new spoken content according to the learning objective and, in certain cases, engage in direct interaction with the student.

4. Areas of Application of AI-Generated Spoken Content in Education

The development of AI-supported voice technologies has made it possible for spoken content to be used in educational settings in a more flexible manner and adapted to diverse learning needs. Unlike conventional audio recordings, AI-generated content can be regenerated according to the learner's level, pace of study, and learning purpose. These features are particularly notable with respect to foreign language teaching, inclusive education, and personalized learning.

4.1. Foreign Language Teaching

One of the most important areas of application for AI-generated spoken content is foreign language teaching. TTS systems enable students to listen to the correct pronunciation of target-language words and sentences, repeat them, and develop their listening skills. In languages such as French, where rhythm, liaison, and the connection between sounds play an important role, supporting written input with spoken examples can contribute to the learning process. Liakin, Cardoso, and Liakina (2017) demonstrated that the use of mobile TTS can serve as a pedagogical tool for developing the production of French liaison.

Another advantage of AI-supported voice systems is the ability to rapidly generate listening materials suited to different proficiency levels. Students can listen to the same content at different speeds and repeat the sections they need. It has also been reported that TTS-supported repeated-reading and listening activities can support fluency and reading-comprehension skills among foreign language learners (Amin, 2022). In addition, generative AI systems can create dialogues tailored to particular communicative situations, allowing students to practice speaking. Spoken content has taken

on a broader pedagogical function, supporting pronunciation, listening, and oral interaction.

4.2. Inclusive Education and Accessibility

AI-supported voice technologies also hold significant potential for students who experience difficulty accessing educational materials. By converting written content into spoken form, TTS systems can facilitate access to digital educational materials, particularly for students with visual impairments. Similarly, presenting text in spoken form for students with reading difficulties or dyslexia can help them focus on the meaning of the content rather than on decoding the text.

Research in this area demonstrates that TTS and similar reading supports can enhance reading-comprehension performance among students with reading difficulties. In their meta-analysis, Wood and colleagues (2018) found that TTS and related read-aloud tools could have a positive effect on the reading-comprehension skills of students with reading difficulties. AI-supported voice systems can therefore be regarded not merely as tools for producing instructional materials, but also as assistive technologies that enhance accessibility in education.

4.3. Personalized Learning

Another important area of application for AI-generated spoken content is the adaptation of learning materials to students' individual needs. Whereas in conventional audio recordings all students follow the same content at the same speed, AI-supported systems can generate texts suited to a student's language level and voice them at different speeds. Thus, a beginner-level student can work with shorter, slower content, while a more advanced student can use longer materials closer to natural speech rate.

Another dimension of personalization is that the pace of repetition and learning can be controlled by the student. Students can relisten to a word or sentence they do not understand as many times as needed and, where necessary, opt for a slower narration. This is particularly important in listening and pronunciation exercises. AI-supported voice systems thus support a shift from a structure in which the student merely consumes a ready-made

audio recording toward a more flexible learning environment in which they can determine their own learning pace and study approach.

Overall, the areas of application of AI-generated spoken content in education can be seen to concentrate on three main points: supporting linguistic and communicative skills in foreign language learning; facilitating access to educational materials for diverse learners; and adapting learning content to individual needs. For these opportunities to be pedagogically effective, however, the accuracy of AI-generated content and its alignment with learning objectives must be evaluated by the teacher.

5. The Pedagogical Potential of AI-Generated Spoken Content

In educational settings, AI-generated spoken content serves a multifaceted pedagogical role, supporting different aspects of the learning process beyond the narration of written materials. In particular, advances in generative AI and speech technologies make it easier to adapt spoken content to a student's level, pace, and learning purpose. Nevertheless, the pedagogical value of these technologies derives not solely from their technical features, but from how they are linked to learning objectives and how they are used within the instructional process.

Within this framework, the pedagogical potential of AI-generated spoken content can be examined across five dimensions: personalized learning, learner autonomy, instant feedback and adaptability, multilingualism and accent diversity, and scalable content production.

5.1. Personalized Learning

One important advantage of AI-supported spoken content is that learning materials can be adapted to students' individual characteristics. In conventional audio recordings, all students listen to the same text, delivered by the same speaker at the same speed. By contrast, AI-supported systems can enable content suited to different proficiency levels and learning needs to be prepared more quickly. This is particularly significant in foreign language teaching, since students within the same classroom may not always share the same level of language proficiency or the same learning needs.

For instance, a text composed of short sentences and basic vocabulary may be voiced at a slower pace for a beginner-level French learner, while a more advanced learner may benefit from more complex syntactic structures voiced at a rate closer to natural speech. The same learning objective can thus be achieved for different students through different inputs. Research on the use of TTS technology in foreign language learning indicates that students can use spoken input for listening, classification, and repetition activities. Notably, Liakin, Cardoso, and Liakina (2017) found that learners who used mobile TTS benefited from the technology in pronunciation practice and concluded that TTS could serve as a supportive tool in second-language pronunciation instruction.

The potential of AI-supported spoken content for personalization is therefore significant in that it allows different materials to be generated without requiring the teacher to prepare a separate audio resource for each student. Personalization, however, should not be understood merely as simplifying a text or altering the speech rate; the content must be linked to the student's actual needs and learning objectives. Artificial intelligence should therefore be regarded not as personalized learning itself, but as a tool that facilitates the creation of personalized learning materials.

5.2. Supporting Learner Autonomy

AI-generated spoken content also holds potential for supporting learner autonomy by increasing students' control over their own learning process. A student's ability to listen to content whenever desired, to repeat sections they do not understand, and to work at their own pace can allow the learning process to become partially independent of the teacher's direct control.

This feature is particularly important in foreign language learning. When a student does not understand a word or sentence, they can relisten to the recording and work with the same material as many times as needed. Amin's (2022) study demonstrated that repeated-reading and listening-while-reading activities conducted through TTS applications could be used to develop the reading fluency and comprehension skills of university-level foreign language students. In this research, students carried out repeated reading and listening activities through TTS applications, and improvements in these skills were observed following the intervention.

Such applications enable students not only to follow the teacher within the classroom, but also to continue their own learning process outside it. It would, however, be inaccurate to equate access to technology directly with learner autonomy. Learner autonomy requires students not only to have access to materials whenever they wish but also to set their own learning goals, select appropriate strategies, and evaluate their own progress. It is therefore important that AI-supported spoken tools be used in a manner that encourages students to make choices within the learning process and to evaluate their own performance, rather than turning them into passive consumers of content.

5.3. Instant Feedback and Adaptability

Another dimension of the pedagogical potential of AI-supported voice technologies is their capacity to provide rapid feedback and adaptive interaction, particularly in foreign language teaching. In conventional audio recordings, students can listen to a model in the target language but cannot receive simultaneous feedback on their own oral production. The integration of speech recognition and AI-supported assessment systems with voice technologies, however, makes it possible for students to receive faster feedback on their own production.

This potential is particularly notable in pronunciation practice. After producing a word or sentence orally, a student can receive feedback from the system and then attempt the same utterance again. The learning process can thus turn into a cycle of production–feedback–reproduction. In a study of EFL students by Li and colleagues (2025), the effect of AI-mediated pronunciation interactions on students' pronunciation performance was examined, and significant improvements were observed among students in the AI-supported group. The study also found that students developed positive perceptions of AI-supported practice.

Adaptability, in turn, extends beyond feedback. Information obtained about a student's performance can enable subsequent activities to be modified in terms of content or difficulty level. For instance, if a student is found to struggle with the production of a particular French sound, spoken activities featuring new words and sentences containing that sound can be prepared. In this way, artificial intelligence ceases to function merely as a system that

presents existing content and becomes a form of learning support capable of generating new content based on the student's performance.

It should not be assumed, however, that feedback provided by AI is in all cases equivalent to human evaluation. In the study by Li and colleagues (2025), while students reported positive views regarding AI-supported tools, some difficulties were also reported concerning the accuracy of speech recognition and feedback. AI-generated feedback therefore appears more appropriately used as a complementary, formative form of support that enables students to practice more frequently, rather than as a substitute for teacher evaluation.

5.4. Multilingualism and Accent Diversity

AI-generated spoken content can also be used to support multilingual learning environments by exposing students to different languages and speech styles. In foreign language teaching in particular, exposing students solely to a single speaker or a single standard voice model may result in the limited representation of the diversity found in real communicative contexts. The ability of AI-supported voice systems to generate content in different languages and with different vocal characteristics allows for the preparation of more diverse auditory input in this respect.

This potential is particularly significant in multilingual educational settings. In classrooms where students use more than one language, spoken content can be used not only to present the target language but also to incorporate students' existing linguistic repertoires into the learning process. The use of voice technologies introduces greater flexibility into the learning process by allowing students to engage with diverse linguistic resources alongside the target language.

A similar potential can be identified concerning accent diversity. Exposing students to different speech styles may prepare them for the linguistic diversity they are likely to encounter in real communicative settings. This potential, however, also raises the question of the extent to which AI systems genuinely represent diverse voices and speech communities. Accent diversity should therefore be assessed not merely as a technical "voice option," but as a matter of linguistic and cultural representation. The following

section will examine this issue in greater detail, focusing on data bias and representation concerns.

5.5. Scalable Content Production

Another pedagogical potential of AI-generated spoken content is the ability to produce a large volume of educational material with diverse characteristics in a shorter period of time. Whereas the preparation of conventional audio materials requires stages such as script writing, narration, recording, and editing, AI-supported systems can substantially accelerate this process, particularly at the narration stage.

This capability can enable teachers to increase the diversity of their materials. For example, short listening texts targeting different language levels on the same topic can be prepared; texts can be revised according to students' needs; and this content can be voiced rapidly. As a result, the diversification of spoken materials, previously limited due to time and production costs, becomes feasible.

Scalability also requires a reconsideration of the teacher's role. Rather than technically producing every material from scratch, teachers may instead select, review, align with pedagogical objectives, and adapt to students' needs the initial drafts generated by AI. The primary contribution of artificial intelligence can therefore be understood as reducing the technical burden of material production, rather than eliminating the teacher's pedagogical responsibility.

However, the acceleration of content production does not automatically mean that the resulting content is of high quality. On the contrary, as production becomes easier, monitoring the accuracy, level-appropriateness, and pedagogical quality of the resulting content becomes even more important. Scalability can therefore become a genuine educational advantage only when combined with human oversight and pedagogical design.

General Assessment

The pedagogical potential of AI-generated spoken content does not derive solely from the technical ease of voice production. The genuine transformation lies in the fact that spoken materials can be adapted to the learner, repeated, rendered interactive, and diversified according to different

learning needs. The use of TTS in pronunciation practice within foreign language teaching (Liakin et al., 2017), its support for repeated-reading and listening activities (Amin, 2022), and the contribution of AI-mediated pronunciation interactions to student performance (Li et al., 2025) illustrate different dimensions of this potential.

A distinction must nevertheless be drawn between pedagogical potential and pedagogical effectiveness. The fact that a technology possesses the capacity for personalization, feedback, or content generation does not imply that it will be effective in every educational setting. In determining the educational value of AI-generated spoken content, attention must therefore be paid not only to technical features but also to dimensions such as linguistic accuracy, pedagogical appropriateness, student needs, cultural representation, and ethical reliability. Indeed, the meta-analysis examining the effect of TTS and similar read-aloud tools on students with reading difficulties shows that these technologies can make a positive contribution to reading comprehension, while also cautioning that this effect may vary depending on contextual and methodological factors (Wood et al., 2018).

The pedagogical potential of AI-generated spoken content should therefore be approached not in terms of the technology replacing the teacher, but as a tool that supports the teacher's pedagogical decisions and expands students' learning opportunities. This approach also provides the basis for the evaluation, undertaken in the following section, of the pedagogical and ethical limitations of these technologies, including vocal naturalness, sociocultural context, data bias, copyright, and voice cloning.

6. Pedagogical and Ethical Limitations

Although AI-generated spoken content offers significant opportunities in educational settings including personalization, accessibility, rapid content production, and interactivity, the educational value of these technologies should not be assumed to arise automatically. Given the emotional, cultural, and contextual functions that voice serves in human communication, artificially generated voices are subject to certain limitations. Furthermore, issues such as data bias, representation, privacy, copyright, voice cloning, and the changing role of the teacher mean that the use of this technology in education cannot be treated merely as a technical matter. UNESCO (2023)

emphasizes that the use of generative AI in education must be approached within a human-centered, safe, equitable, and meaningful framework, drawing particular attention to data privacy, ethical validation, and pedagogical design.

For this reason, technical criteria such as "how natural it sounds" or "how quickly it is produced" are not sufficient when evaluating AI-generated spoken content. For a voice to be pedagogically meaningful, it must also be capable of appropriately conveying emotion, context, cultural meaning, and communicative purpose. These limitations are addressed under the headings below.

6.1. Vocal Naturalness and the Emotional Dimension

Although current text-to-speech systems have come to closely approximate human voices, vocal naturalness is not equivalent to possessing all the qualities of human communication. Human speech does not consist solely of the correct articulation of words; prosodic features such as pitch, rhythm, stress, pausing, speech rate, and intonation play an important role in conveying a speaker's emotional and communicative intent. Consequently, even when an artificial voice is linguistically accurate and fluent, it may not convey context-dependent emotional meaning with the same degree of accuracy.

Recent studies also draw attention to this issue. Ye and Yang (2026) identify an "emotional gap" in AI-supported speech synthesis, particularly in its ability to convey the complexity of human emotions. They note that complex emotions, contextual nuances, and prosodic features may not be adequately represented in artificial voices. Similarly, a study examining the emotional transmission of computer-generated voices in educational multimedia lessons found that although artificial voices could convey certain basic emotions, these emotions did not always have the expected effect on students' social connection and learning. This finding indicates that voice does not function merely as a channel for conveying information but also plays a role in the emotional and social dimensions of the learning experience.

This limitation is particularly important in foreign language teaching. Students must learn not only how words are pronounced in the target language, but also how an utterance is delivered. The same sentence, for example, can

be intoned differently to express surprise, anger, politeness, irony, or doubt. The accurate production of words by an artificial voice does not, on its own, guarantee that a student will correctly grasp the communicative meaning.

For this reason, the evaluation of spoken content for educational purposes should consider not only technical naturalness but also prosodic and emotional appropriateness. In situations including communicative skills, drama activities, or tasks requiring emotional expression in particular, replacing the teacher's authentic human voice and face-to-face interaction entirely with an artificial voice may create pedagogical difficulties.

6.2. The Lack of Pragmatic and Sociocultural Context

Language learning does not consist solely of producing grammatically correct sentences. The context in which an utterance is made, to whom, for what purpose, and in what tone can substantially alter the meaning of communication. Pragmatic competence is therefore as important as linguistic accuracy in foreign language education. One of the significant limitations of AI-generated spoken content is that such contextual and sociocultural features are not always adequately represented.

Humor, irony, implied meaning, politeness strategies, and cultural references, for example, cannot be understood solely on the basis of word meaning; the same utterance may be interpreted differently across cultural contexts. A comprehensive systematic review of the current EFL/ESL literature reveals that dimensions such as pragmatic nuance, cultural appropriateness, humor, empathy, and linguistic diversity continue to pose significant challenges for AI-supported language-learning tools.

This is particularly important when AI is used to generate dialogues for foreign language teaching. A system may produce a grammatically correct dialogue yet fail to recognize that a particular utterance could be perceived as rude, overly formal, ironic, or otherwise inappropriate in a given cultural context. In such cases, students may not only learn incorrect linguistic forms but also internalize inappropriate ways of communicating as acceptable.

A distinction must therefore be drawn, in AI-generated spoken content, between "correct French," "correct English," or "correct expression" in another language, and "contextually appropriate communication." The teacher's role should extend beyond checking linguistic accuracy to

evaluating the sociocultural and pragmatic appropriateness of the content. Therefore, AI-generated dialogues should be reviewed by the teacher before being used in foreign language teaching.

6.3. Data Bias and Representation Problems

Another important limitation of AI systems is that the voices and language models used in education may represent certain communities more than others. AI models are trained on large datasets that may contain existing inequalities, which can consequently be reflected in their outputs. This issue is particularly relevant to accent, age, gender, and low-resource languages.

Research on speech technologies indicates that automatic speech-recognition systems may not perform equally well across different groups of speakers. The systematic review by Feng and colleagues (2024) shows that automatic speech-recognition systems perform differently depending on gender, age, regional accent, and the speaker's native language. The researchers note that this bias cannot be explained by pronunciation differences alone and that solutions specific to language and model architecture are needed.

This issue can have direct pedagogical consequences for foreign language education. If an AI system recognizes a particular accent more readily than others, it may judge a student's speech to be of lower quality even when it is intelligible to a human listener. Indeed, the study by Bae (2026) found that AI-based speech-recognition systems exhibited differing levels of accuracy in recognizing the speech of English speakers with different native-language backgrounds. Similar representation issues may also arise in synthetic voices.

Research conducted by Michel and colleagues (2025) on commercial synthetic voice services such as Speechify and ElevenLabs shows that technical performance differences exist across different English accents, and that certain vocal characteristics may reinforce existing linguistic inequalities.

For this reason, "accent diversity" should not be measured merely by the number of different voice options available within a system. The more important question is which linguistic and speech communities are represented, and which remain invisible. This problem may be particularly pronounced for low-resource languages and underrepresented accents. In such cases, AI may reinforce existing linguistic hierarchies rather than promote diversity in education.

6.4. Copyright and Voice-Cloning Problems

With the advancement of AI-supported voice technologies, it has become possible to reproduce a person's voice and to artificially imitate a specific individual's vocal characteristics. While this technology offers educational opportunities, such as allowing teachers to create materials using their own voices, it also raises concerns about consent, privacy, digital identity, copyright, and misuse.

One of the fundamental concerns regarding voice cloning is that a voice should not be regarded merely as a "voice file." Since a person's voice is closely linked to their identity, copying or using it without permission may raise concerns about identity and privacy. McGettigan and Bloch (2023) note that voice-cloning technologies may have psychological and ethical consequences because of the personal and identity-related nature of the human voice.

Similarly, Fenwick and colleagues (2024) argue that voice-cloning technologies create new legal challenges that extend beyond existing frameworks of copyright, privacy, and personality rights, and that an individual's control over their own voice, together with explicit consent, should be treated as fundamental principles. Using a cloned voice without the individual's knowledge in fabricated statements, advertisements, or misleading educational materials may further raise concerns about reliability.

A similar situation may arise in educational settings. A teacher's voice, for example, could be cloned without permission and used in instructional materials, or a statement the teacher never made could be generated in their voice and presented as authentic. For this reason, the use of voice cloning should follow principles of explicit consent, clearly defined purposes and duration of use, data retention, and transparent disclosure of artificially generated content. UNESCO likewise places particular emphasis on safeguarding data privacy, intellectual property, and a human-centered ethical approach in the use of generative AI in education (UNESCO, 2023)

Although cloning a teacher's voice may be pedagogically useful, its technical feasibility does not necessarily make it ethically acceptable.

6.5. The Changing Role of the Teacher

The growing prevalence of AI-generated spoken content has also raised important questions about the teacher's role in the educational process. The preparation of spoken materials, their adaptation to different levels, and, in some cases, the provision of automated feedback to students demonstrate that certain tasks traditionally carried out by teachers can now be supported by technological tools. This should not be seen as replacing the teacher, but rather as a transformation of the teacher's role.

The teacher's role is therefore shifting from simply delivering content or preparing spoken materials to selecting, verifying, and aligning AI-generated content with pedagogical objectives, while guiding students in its use. UNESCO's guidance on generative AI likewise emphasizes that rather than replacing human intelligence, AI should be adopted within a human-centered approach that supports both teachers and students (UNESCO, 2023).

This transformation is particularly important for AI-based voice tools, since even if a system is technically capable of producing correct pronunciation, it cannot, on its own, determine which voice should be presented to which student, for what purpose, and in what context. Teachers can assess the appropriateness of AI-generated material by taking into account students' prior knowledge, motivation, communicative needs, and cultural context.

The spread of AI in education, therefore, need not diminish the teacher's importance; instead, it may make the teacher's responsibility for pedagogical decision-making more visible. The teacher's fundamental role lies not in increasing the volume of content that technology can generate, but in determining under what conditions such content can be converted into learning value. UNESCO's AI competency framework for teachers likewise supports this transformation by identifying skills such as ethical use, data privacy, intellectual property, and the evaluation of AI-generated content as part of teachers' core competencies (UNESCO, 2024).

General Assessment

The limitations of AI-generated spoken content demonstrate that this technology cannot be evaluated solely on the basis of technical performance. The naturalness of a voice does not imply that it is emotionally and

pragmatically adequate; support for diverse accents does not imply that a system is fair and representative; and the technical capacity to clone a person's voice does not imply that doing so is ethically or legally legitimate.

For this reason, when evaluating AI-supported voice technologies for use in education, technical quality must be distinguished from pedagogical and ethical appropriateness. In foreign language teaching in particular, linguistic accuracy, pragmatic appropriateness, accent diversity, and cultural representation must be considered together. Likewise, it is important that teachers not only be able to use these technologies, but also be able to critically evaluate the content they produce. In this respect, the role of artificial intelligence in education should be conceived not as an automatic content generator, but as a pedagogical tool guided and supervised by human judgment.

This assessment provides the basis for the "Analytical Framework for Evaluating AI-Based Voice Resources in Education" presented in the following section, since the limitations discussed above indicate that a given voice-based AI resource must be assessed not only in terms of its technical performance but also with regard to linguistic and pedagogical appropriateness, accessibility, sociocultural considerations, and ethical issues.

7. An Analytical Framework for Evaluating AI-Based Voice Resources in Education

Assessments concerning the use of AI-generated spoken content in educational settings indicate that these technologies cannot be considered solely in terms of their technical performance. As demonstrated in the preceding sections, AI-supported voice technologies offer significant opportunities for personalized learning, learner autonomy, accessibility, and scalable content production. However, they also have various limitations related to the emotional dimension of voice, pragmatic and sociocultural context, data bias, representation, copyright, and voice cloning. Therefore, when evaluating a voice-based AI resource, it is insufficient for educators to focus solely on questions such as "Does the voice sound natural?" or "Does the system function correctly?"

On this basis, an analytical framework consisting of five core dimensions (linguistic quality, pedagogical appropriateness, accessibility, ethical reliability, and technical quality) can be proposed for evaluating AI-supported voice resources intended for educational use. The purpose of this framework is not to declare a given AI tool superior to another, but to help educators evaluate different voice technologies in a systematic and critical manner. This approach also aligns with UNESCO's view that generative AI in education should be human-centered, safe, and aligned with pedagogical purposes (UNESCO, 2023).

7.1. Linguistic Quality

The first dimension for evaluating a voice-based AI resource is linguistic quality. It is not sufficient for the generated voice merely to be intelligible; pronunciation, stress and intonation, vocabulary, and grammatical structures must also be appropriate for the target language. Given that, in foreign language teaching in particular, students may use an artificially generated voice as a model, the quality of the linguistic input produced by the system carries direct pedagogical significance.

In this respect, the following questions may be posed:

- Is pronunciation accurate and intelligible with respect to the target language?
- Are word-stress patterns appropriate?
- Is the speech rate appropriate for the student's level?
- Do intonation and pausing support meaning?
- Is the system capable of producing different speech styles with sufficient accuracy?

As noted earlier, TTS technologies can be used in pronunciation practice in foreign language teaching in particular. Liakin, Cardoso, and Liakina (2017) demonstrated that mobile TTS applications can be used in teaching French liaison. However, the fact that a given technology can produce a particular pronunciation feature does not mean that it will be equally successful across all linguistic contexts.

Educators should therefore treat AI-generated voice not as a "correct model" to be accepted uncritically, but as linguistic input whose accuracy must be verified.

7.2. Pedagogical Appropriateness

The second dimension concerns the extent to which the spoken content aligns with a specific learning objective. A technically high-quality voice may not be pedagogically appropriate. For example, using an excessively slow and artificial-sounding voice in a listening activity designed for advanced learners may limit their preparation for real communicative settings. Similarly, overly fast or complex spoken content intended for beginner-level students may increase cognitive load.

The following questions therefore become important:

- Does the spoken content align with the specified learning outcome?
- Is the language level of the content appropriate to students' proficiency level?
- Does the content support students' active engagement?
- Does the voice genuinely contribute to the pedagogical purpose of the activity?
- Does the use of AI provide a meaningful contribution to the learning process?

This last question is particularly important. The fact that content can be generated using AI does not necessarily mean that it should be. If the same learning objective can be achieved more effectively through a conventional material prepared by the teacher, the pedagogical rationale for using AI should be reconsidered.

Amin's (2022) study demonstrated that the use of TTS in repeated-reading and listening activities can support students' fluency and comprehension skills. This suggests that the pedagogical value of TTS should be assessed not merely in terms of its capacity to "produce voice," but in relation to the learning activity in which the student engages.

7.3. Accessibility

The third dimension concerns whether AI-generated spoken content supports the participation of diverse groups of students in the learning process. Voice technologies can offer important opportunities, particularly for students with visual impairments, students with reading difficulties, and individuals with other learning needs.

The meta-analysis by Wood and colleagues (2018), examining the effect of TTS and similar audio-reading technologies on students with reading difficulties, demonstrates that such technologies can be supportive in reading-comprehension processes. AI-supported spoken content can therefore provide an alternative or complementary channel of access to written materials.

Accessibility, however, should not be understood merely as "the narration of text." Attention should also be given to whether students can adjust the speech rate, relisten to content, switch between text and audio, and, where necessary, obtain written support.

In this context, the relevant evaluation questions are as follows:

Can the content be used by students with diverse learning needs?

Are options for speech rate and repetition available?

Can spoken and written content be used together?

Is the user interface accessible?

Does the technology exclude any particular group of students?

Accessibility should therefore be regarded not as an additional feature of the technology, but as one of the fundamental evaluation criteria for AI resources used in education.

7.4. Ethical Reliability

The fourth dimension concerns whether AI-generated spoken content is ethically reliable. As discussed in the previous section, issues such as voice cloning, data use, the protection of personal data, copyright, and consent are also important in educational settings. The unauthorized use of a teacher's or another individual's voice constitutes not merely a technical concern, but also an ethical and legal problem.

Educators should therefore ask the following questions:

- Has explicit consent been obtained for the voice used?
- How was the voice data obtained, and how is it stored?
- Is the user informed that the content has been artificially generated?
- Is the copyright status of the materials used clear?
- Does the generated voice imitate a person's identity without permission?
- Does the system produce biases directed toward particular languages or communities?

UNESCO (2023) emphasizes that the use of generative AI tools in education must take into account data privacy, intellectual property, security, and a human-centered approach. Similarly, UNESCO's AI competency framework for teachers indicates that teachers need not only the ability to use AI systems on a technical level, but also the ability to evaluate them ethically and critically (UNESCO, 2024).

Ethical reliability should therefore be evaluated separately from the other dimensions. A system that functions technically well or appears pedagogically useful should not be used directly in an educational setting if it raises ethical concerns.

7.5. Technical Quality

The final dimension concerns the technical performance of the system. This includes the intelligibility, fluency, latency, voice quality, and cross-device functionality of the generated speech. Although technical quality is necessary for the other dimensions to be realized, it is not sufficient on its own.

For example, a system that produces a highly natural-sounding voice cannot be considered a pedagogically reliable resource if it generates mispronunciations or culturally inappropriate content. Similarly, a system with high technical quality may still have limitations in terms of accessibility.

The core questions regarding technical quality are therefore as follows:

- Is the voice intelligible and consistent?

- Are natural speech characteristics adequately reproduced?
- Does the system function reliably across different speeds and voices?
- Do technical errors occur during generation?
- Is the system easy and sustainable for users to operate?

The fundamental principle here is that technical performance must be evaluated alongside the other four dimensions.

7.6. Proposed Evaluation Model

Taking these five dimensions together, the following model can be proposed for evaluating AI-generated spoken content in education:

Dimension	**Core Evaluation Questions**
Linguistic Quality	Is pronunciation accurate? Are stress and intonation appropriate?
Pedagogical Appropriateness	Does it align with the target outcome and student level?
Accessibility	Can it be used by diverse groups of students?
Ethical Reliability	Is it safe in terms of data, consent, copyright, and representation?
Technical Quality	Is the voice intelligible, natural, and stable?

The defining feature of this framework is that the dimensions complement one another. A voice does not need to perform perfectly across every dimension to be useful in education. However, strong technical performance cannot compensate for serious ethical or pedagogical problems. In other words, before asking whether an AI-generated voice is "good," one must first ask for what purpose, with which group of students, and within what educational context it will be used.

This approach makes it possible to evaluate AI-supported voice technologies in a manner centered on pedagogy rather than on the tool itself.

8. Future Directions and Concluding Perspectives

The development of AI-generated spoken content in education does not appear to be limited merely to the production of increasingly natural-sounding voices. Future developments are likely to focus on making voice technologies more responsive to the learning process, enabling them to understand diverse communicative contexts and engage in real-time interaction with students. This transformation holds considerable potential, particularly for foreign language education.

First, the development of emotional speech synthesis can be expected. Although current systems can vary stress, pitch, and speech rate to a certain degree, they do not yet fully capture the emotional and contextual diversity of human speech. In the future, voice-generation systems may become capable of modeling not only "what is said" but also, more successfully, "how it is said." Such a development could offer new opportunities in foreign language teaching for developing pragmatic competence, role-play, and communicative skills. At the same time, it should be borne in mind that this progress may also raise concerns regarding cultural representation and emotional artificiality, as discussed in earlier sections.

A second important area of development will be real-time personalization. The instantaneous adjustment of spoken content according to a student's speech rate, error types, or performance may create a learning experience distinct from today's static audio materials. If a student does not understand an utterance, the system could speak more slowly; if a student struggles with a particular word, the same word could be re-presented in different contexts. Spoken content could thus evolve from a static, pre-prepared material into a learning tool continuously adapted to the student's performance.

Third, AI-supported voice technologies can be expected to be used more widely in multilingual educational settings. In classrooms where more than one language is used within the same educational context, the ability of systems to switch between different languages may allow students' linguistic repertoires to be more fully incorporated into the learning process. For this potential to be realized, however, low-resource languages and diverse accents must be adequately represented within technological systems.

Otherwise, rather than supporting multilingualism, AI may instead further reinforce dominant languages and standard linguistic forms.

Another important dimension of the future is human–AI collaboration. Rather than evaluating the development of AI in education in terms of the teacher's disappearance, restructuring the division of tasks between the teacher and AI represents a more realistic approach. AI can generate large volumes of content, provide narration, and enable students to engage in repetition; however, determining the educational purpose of content, assessing its cultural appropriateness, and interpreting students' needs still require substantial human expertise. UNESCO's AI competency framework for teachers likewise emphasizes the importance of teachers' capacity for ethical, pedagogical, and critical decision-making in the age of artificial intelligence (UNESCO, 2024).

For this reason, rather than the teacher's role disappearing in the future, it can be expected to be redefined. The teacher may move from being a simple user of AI-generated content toward becoming a designer responsible for selecting, verifying, adapting, and situating such content within a pedagogical context. In such an approach, AI provides generative capacity, while the teacher provides pedagogical and ethical decision-making capacity.

9. Conclusion

AI-generated spoken content constitutes a significant area of transformation in the development of educational technologies. The trajectory extending from text-to-speech systems to voice cloning and to speech systems integrated with large language models has carried the use of voice in education beyond conventional audio recordings. Today, spoken content is evolving from ready-made materials that students simply listen to into resources that can be personalized, regenerated, and made interactive to a certain degree.

This transformation is particularly noteworthy with respect to foreign language teaching. The usefulness of TTS technologies for pronunciation and listening practice (Liakin et al., 2017; Amin, 2022), the contribution of AI-supported interactions to pronunciation practice (Li et al., 2025), and the capacity of voice technologies to provide access for learners with diverse

needs collectively demonstrate the considerable pedagogical potential of these technologies. However, realizing this potential depends not only on the technical capabilities of the technology but also on how effectively it is integrated into pedagogical practice.

The limitations addressed in this chapter reinforce this point. The potential inadequacy of artificial voices in emotional and pragmatic terms, biases in data and representation, inequalities in accent diversity, concerns regarding consent and privacy in voice cloning, and the changing role of the teacher all necessitate a critical evaluation of AI use in education. The adoption of AI-generated spoken content in education should therefore be considered not simply in terms of whether the technology should be used, but in terms of the educational purposes it serves, the conditions under which it is used, and the mechanisms needed to ensure appropriate oversight.

The five-dimensional analytical framework proposed in this respect (linguistic quality, pedagogical appropriateness, accessibility, ethical reliability, and technical quality) can help educators systematically evaluate different voice-based AI resources. The core premise of this framework is that technological innovation should not be taken as a direct indicator of pedagogical value. Even the most advanced AI system has educational value only when it is aligned with students' needs, learning objectives, cultural contexts, and ethical principles.

In conclusion, the future of AI-generated spoken content should be sought not in artificial systems that entirely replace the human voice, but in a complementary relationship between human beings and artificial intelligence. AI can accelerate content production, generate materials suited to diverse learning needs, and support students' individual practice. The teacher, by contrast, will remain the central agent responsible for ensuring that such content is rendered pedagogically, culturally, and ethically meaningful. For this reason, the key issue for the future of education is not whether artificial intelligence should be used, but how it can be designed and implemented to serve educational purposes.

METHOD NOTE: DECLARATION OF AI USE

In preparing this chapter, the AI language model developed by Anthropic, specifically the Sonnet 5 version, was accessed through the claude.ai interface. The AI system was used for three delimited purposes: (1) an academic-quality review of the manuscript and a limited scientific and bibliographic editorial pass, including verification of the match between in-text citations and the reference list and the identification and completion of missing citations with verifiable sources; (2) grammatical and stylistic refinement in the process of translating the manuscript into academic English; and (3) formatting of the document in Word (.docx). All sources, citations, arguments, and the analytical framework presented in this chapter belong to the authors; the role of the AI system was limited to editorial support, citation verification, translation assistance, and document formatting, all of which were carried out under human supervision. The manuscript was reviewed and approved by both authors prior to submission.

References

Abdous, M., Facer, B. R., & Yen, C.-J. (2012). Academic effectiveness of podcasting: A comparative study of integrated versus supplemental use of podcasting in second language classes. Computers & Education, 58(1), 43–52.

Alm, A. (2013). Extensive listening 2.0 with foreign language podcasts. Innovation in Language Learning and Teaching, 7(3), 266–280.

Amin, E. A.-R. (2022). Using repeated-reading and listening-while-reading via text-to-speech apps in developing fluency and comprehension. World Journal of English Language, 12(1), 211–220.

Autio, C., Schwartz, R., Dunietz, J., Jain, S., Stanley, M., Tabassi, E., Hall, P., & Roberts, K. (2024). Artificial intelligence risk management framework: Generative artificial intelligence profile (NIST AI 600-1). National Institute of Standards and Technology.

Bae, Y. (2026). Investigating accent bias in AI intelligibility of L2 English accents. TESOL Quarterly.

Bone, E. K., & Bouck, E. C. (2017). Accessible text-to-speech options for students who struggle with reading. Preventing School Failure: Alternative Education for Children and Youth, 61(1), 48–55.

Feng, S., Halpern, B. M., Kudina, O., & Scharenborg, O. (2024). Towards inclusive automatic speech recognition. Computer Speech & Language, 84, 101567.

Fenwick, M., Jurcys, P., Liaudanskas, A., Earp, B. D., Porsdam Mann, S., & Kozuka, S. (2024). Voice cloning in an age of generative AI: Mapping the limits of the law & principles for a new social contract with technology. SSRN.

Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2024). Generative AI. Business & Information Systems Engineering, 66(1), 111–126.

Handley, Z. (2009). Is text-to-speech synthesis ready for use in computer-assisted language learning? Speech Communication, 51(10), 906–919.

Kitao, K. (1995). The history of language laboratories: Origin and establishment. ERIC.

Klatt, D. H. (1987). Review of text-to-speech conversion for English. The Journal of the Acoustical Society of America, 82(3), 737–793.

Li, L., Zhang, X., Zou, B., & Yang, Q. (2025). AI partner or peer partner? Exploring AI-mediated interaction in EFL pronunciation from a socio-cultural perspective. Learning, Culture and Social Interaction, 55, 100958.

Li, M., Ahmadiadli, Y., & Zhang, X.-P. (2025). A survey on speech deepfake detection. ACM Computing Surveys, 57(7), Article 165.

Liakin, D., Cardoso, W., & Liakina, N. (2017). The pedagogical use of mobile speech synthesis (TTS): Focus on French liaison. Computer Assisted Language Learning, 30(3–4), 325–342.

McGettigan, C., Bloch, S., Bowles, C., Dinkar, T., Lavan, N., Reus, J. C., & Rosi, V. (2023). Voice conversion and cloning: Psychological and ethical implications of intentionally synthesising familiar voice identities. *Journal of the British Academy, 13*(3), Article a31.

Michel, S., Kaur, S., Gillespie, S. E., Gleason, J., Wilson, C., & Ghosh, A. (2025). "It's not a representation of me": Examining accent bias and digital exclusion in synthetic AI voice services. arXiv.

Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T.-Y. (2021). FastSpeech 2: Fast and high-quality end-to-end text to speech. International Conference on Learning Representations (ICLR).

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R. J., Saurous, R. A., Agiomyrgiannis, Y., & Wu, Y. (2018). Natural TTS synthesis by conditioning

WaveNet on mel spectrogram predictions. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 4779–4783.

T.C. Cumhurbaşkanlığı Dijital Dönüşüm Ofisi Başkanlığı, & T.C. Sanayi ve Teknoloji Bakanlığı. (2021). Ulusal Yapay Zekâ Stratejisi 2021–2025. Ankara.

UNESCO. (2023). Guidance for generative AI in education and research. UNESCO.

UNESCO. (2024). AI competency framework for teachers. UNESCO.

Vanderplank, R. (2010). Déjà vu? A decade of research on language laboratories, television and video in language learning. Language Teaching, 43(1), 1–37.

Wood, S. G., Moxley, J. H., Tighe, E. L., & Wagner, R. K. (2018). Does use of text-to-speech and related read-aloud tools improve reading comprehension for students with reading disabilities? A meta-analysis. Journal of Learning Disabilities, 51(1), 73–84.

Xiong, Y. (2026). Optimization of foreign language listening teaching strategies based on intelligent TTS. Discover Artificial Intelligence.

Yang, Z., Shimizu, S., Yu, Y., & Chu, C. (2025). When large language models meet speech: A survey on integration approaches. Findings of the Association for Computational Linguistics: ACL 2025, 20298–20315.

Ye, J., & Yang, J. (2026). Bridging the Affective Gap: A Pedagogical Framework for Critical Voice Artistry in the AI Era. *Journal of Educational Technology and Innovation,* 8(1), 12–24.

Adaptive Intelligence and the Mediating Teacher: A Critical-Pedagogical Account of Artificial Intelligence and Adaptive Learning Management Systems in Foreign Language Education

Gökhan Dinar[1] & *Ali Tilbe*[2]

1. Introduction: The Digital Turn and the AI Paradigm in L2 Acquisition

Few human capacities have been as continuously reshaped by their delivery technologies as the learning of a foreign language. From the grammar-translation tradition through the audio-lingual laboratory to the communicative classroom, methods of teaching, materials design, and assessment have evolved alongside the media that carried them. For most of this history, however, instruction rested on a structurally similar premise: a teacher operated as the principal source of knowledge, a standardised syllabus was delivered homogeneously to a class, and individual cognitive differences were, for practical reasons, largely set aside. This one-size-fits-all logic was never a pedagogical ideal so much as a constraint of scale.

The integration of information and communication technologies into education, and particularly the rapid maturation of artificial intelligence over

[1] Asst. Prof. Dr., Istanbul University Cerrahpaşa, Hasan Âli Yücel Faculty of Education, Department of Foreign Language Education, Division of French Language Education, AYG-GLOBAL Education Technologies and Software Company, ORCID: 0000-0002-6057-7042, *gokhan.dinar@iuc.edu.tr*, Türkiye

[2] Prof. Dr., Istanbul University Cerrahpaşa, Hasan Âli Yücel Faculty of Education, Department of Foreign Language Education, Division of French Language Education, AYG-GLOBAL Education Technologies and Software Company, ORCID: 0000-0003-4634-8822, *ali.tilbe@iuc.edu.tr*, Türkiye

the past decade, has unsettled that premise. AI-supported platforms, and the Adaptive Learning Management Systems (ALMS) that increasingly form their pedagogical backbone, promise to replace static transmission with dynamic, data-informed environments that reshape themselves in response to a learner's performance. The shift is often described as a move from Computer-Assisted Language Learning (CALL) to Intelligent CALL, or AI-mediated language learning, in which natural language processing (NLP), machine learning, large language models (LLMs), and automatic speech recognition (ASR) operate in coordination with language-acquisition objectives.

Rather than ask whether AI works in language teaching, a question too coarse to be answered, this chapter asks where these technologies produce genuine learning gains, for which competences, and at what epistemic and ethical cost. It proceeds as a critical synthesis of recent meta-analytic and systematic-review evidence, read through a sociocritical and mediation-oriented lens. Three commitments organise the argument. The first is methodological caution: several claims that circulate freely in the applied literature, including the durable myth that instruction should be matched to a learner's learning style, do not survive contact with the evidence, and we say so plainly. The second is critical: adaptive systems present themselves as neutral instruments, yet the corpora on which they are trained carry the imprint of a language's dominant varieties, and it is the speakers at the margins who absorb the cost of that inheritance. The third is constructive: drawing on the Common European Framework of Reference (Council of Europe, 2020), we argue that the responsible horizon is neither full automation nor technophobic refusal but a hybrid pedagogy in which the teacher is repositioned as a mediator and AI itself becomes an object of critical scrutiny.

2. The Technological and Pedagogical Ecosystem of AI-Supported Platforms

The integration of AI into language teaching extends well beyond the delivery of materials. At its most ambitious it models aspects of the acquisition process itself and optimises the interaction between learner and content. This matters because L2 learning is a complex neurocognitive undertaking: lexis, syntax, phonology, comprehension, and pragmatics constitute

distinct yet tightly coupled subsystems that must be developed simultaneously. AI-based platforms increasingly offer specialised analytic instruments for each of these strands, lightening the teacher's managerial load while allowing learners to progress at an individual pace.

Pronunciation and speaking illustrate the affordance most clearly. ASR engines coupled with phonetic analysis compare a learner's realisation of a phoneme, its stress and intonation, against reference acoustic values and return corrective feedback within seconds. The pedagogical value here is real and structural: in crowded classrooms the speaking time available to any individual learner is minimal, and many learners avoid oral practice altogether because of foreign language anxiety. By offering a private, non-judgemental space for unlimited rehearsal, such systems lower the affective filter and create conditions for the kind of repeated, feedback-rich practice that the classroom cannot supply. As later sections argue, however, the same acoustic modelling that makes this possible is also the locus of one of the field's most serious equity problems.

In writing, NLP-based tools for text generation, translation, and grammatical checking now operate at a level far beyond surface spell-checking. Drawing on large language models, they assess syntactic accuracy, semantic coherence, and stylistic register, and they can indicate not only that a formulation is non-target-like but also why, and how it might be rendered more idiomatically. The quantitative evidence for generative AI is also beginning to consolidate. In the first meta-analysis devoted to generative AI chatbots in second language acquisition, Li, Wang, and Yang (2025) synthesised 41 experimental and quasi-experimental studies (48 effect sizes; 3,515 participants) and found a moderate-to-large overall effect on language development (effect size = 0.576, 95% CI [0.385, 0.768]), with the clearest gains in vocabulary learning and in target languages other than English. The corollary for teachers is significant: part of the time-intensive labour of correcting low-level error can be delegated, freeing instructors to invest in higher-order interventions.

Listening and the design of aural input have been transformed by neural text-to-speech (TTS) systems, which can render written content in multiple languages, regional accents, age registers, and emotional tones. Where learners once encountered only their teacher's accent and a limited set of

textbook recordings, synthetic voices multiply exposure to the variation that characterises authentic communication. Used judiciously, these tools enrich aural input and diversify the channels available to learners.

As the ecosystem matures, the teacher's role is itself transformed. The familiar slogan that the instructor moves from *sage on the stage* to *guide on the side* captures part of this shift but understates it; the more precise and more demanding description, to which Section 8 returns, is that of mediator. A mature ALMS functions as the nerve centre of this ecosystem, pooling data from heterogeneous tools and returning insights on which someone must act. Whether those insights are pedagogically valid, and whether their statistical basis is transparent to the humans acting on them, are open questions taken up below.

3. Architecture and Mechanisms of Adaptive Learning Management Systems

An Adaptive Learning Management System, by contrast to conventional LMS repositories, interprets each interaction between learner and content algorithmically and reacts to the learner's cognitive state. The design goal is to maximise both the efficiency and the effectiveness of instruction, where efficiency denotes reaching target outcomes with the least wasted time given the learner's existing knowledge and preferences.

Technically, an ALMS typically rests on a multi-tier architecture. A presentation layer delivers a responsive interface across web and mobile. A data layer, built on relational or NoSQL databases, records demographic information alongside every click, test result, session duration, and erroneous selection. An application layer houses the system's intelligence: machine-learning engines and, in current systems, LLMs integrated through application programming interfaces, which allow learner responses to be evaluated with semantic depth and which can generate personalised content on demand.

Functionally, adaptation rests on four pedagogical components: adaptive content, adaptive sequencing, adaptive assessment, and real-time feedback. Through an initial diagnostic, or through accumulated interaction data, the system constructs a proficiency map of the learner. Adaptive content filters materials to the learner's current level, while adaptive sequencing reorders

topics according to the mastery of prerequisite concepts. Adaptive assessment, in its computerised adaptive testing (CAT) form, modifies item difficulty in response to each answer. Real-time, guidance-oriented feedback after each attempt allows the learner to detect and repair errors before they consolidate.

This machinery has a respectable theoretical pedigree. The principle of holding a learner just beyond current competence echoes Vygotsky's (1978) zone of proximal development, and the consolidation of knowledge through spaced, response-sensitive practice is supported by research on adaptive scheduling (Mettler, Massey, & Kellman, 2011). A methodological correction is nonetheless required, because a popular and superficially similar idea must be firmly distinguished from the one just described.

3.1 A Necessary Correction: Adaptation to Prior Knowledge, Not to Learning Styles

Much of the applied and design literature treats adaptation to a learner's learning style (typically the visual, auditory, or kinesthetic preferences of instruments such as the Felder-Silverman model) as if it were an evidence-based personalisation strategy. It is not. The relevant empirical claim, known as the meshing hypothesis, holds that learners taught in their preferred modality will learn more than those taught otherwise. Reviewing the literature against the experimental design that would be required to test it, Pashler, McDaniel, Rohrer, and Bjork (2008) concluded that there was no adequate evidence to justify incorporating learning-styles assessment into general educational practice. Later commentators have seen no reason to revise that verdict (Riener & Willingham, 2010; Kirschner, 2017).

The distinction is not pedantic. Adaptation to prior knowledge and demonstrated mastery rests on solid ground; adaptation to putative perceptual style does not. A scientifically responsible ALMS should ground personalisation in prior knowledge, mastery, error history, and goals, and should treat learning-style matching as, at best, an unproven heuristic. With that correction in place, the embedding of LLMs within adaptive systems represents a genuine advance, enabling the system to handle translation, complex natural-language understanding, and generation, and to adapt the topic of materials to learner interest in ways that have a credible motivational rationale.

4. Empirical Effects on Language Skills and Self-Regulated Learning

Beyond architecture lies the question of effect, and here the strongest evidence is not the single striking experiment but the accumulated weight of meta-analysis. For intelligent tutoring systems (ITS), the broad ancestor of adaptive language platforms, VanLehn's (2011) influential review found that ITS raised outcomes over conventional instruction with an effect size of $d = 0.76$, strikingly close to the $d = 0.79$ he estimated for human tutoring. Ma, Adesope, Nesbit, and Liu (2014), analysing 107 effect sizes from 14,321 participants, reported that ITS outperformed teacher-led large-group instruction ($g = 0.42$) and non-ITS computer-based instruction ($g = 0.57$), with no significant difference from one-to-one human tutoring. Kulik and Fletcher (2016) reached a similarly encouraging verdict, reporting a median effect of 0.66 across fifty controlled evaluations, while Steenbergen-Hu and Cooper (2014), who focused on college students, found more modest but still positive gains ($g \approx .32–.37$). The spread is itself instructive: the size of the advantage depends on the population studied, the outcome measure, and the comparison condition. One caveat belongs on the table before any of these figures are borrowed for language education. The ITS corpus is dominated by mathematics, science, and programming, so treating its effect sizes as portable to L2 learning is an assumption rather than a finding, and one more reason for the methodological caution this chapter urges throughout.

With that reservation registered, the reasonable summary is that well-designed adaptive systems reliably outperform undifferentiated whole-class instruction while approaching, though not exceeding, the gold standard of individual human tutoring. Figures of this kind should replace the inflated, unsourced percentages that circulate in promotional accounts. The gain is located where mechanism predicts it: in the immediate, individualised correction of rule-governed error. When a syntactic or morphological slip is flagged in milliseconds and the learner re-attempts the same structure in a new context, fossilisation is forestalled. It is on accuracy, pronunciation, and lexical retrieval that adaptive systems show their clearest advantage, a pattern consistent with the meta-analytic finding that generative AI

chatbots benefit L2 development most visibly in vocabulary learning (Li, Wang, & Yang, 2025).

The motivational and self-regulatory dimension deserves equal attention, for it is in the affective domain that the one-size-fits-all model fails most visibly. By holding the learner in a productive zone, neither bored nor overwhelmed, and by supplying continuous diagnostic feedback, adaptive platforms can support the development of self-regulated learning (SRL): the capacity to set goals, select strategies, monitor progress, and self-evaluate (Zimmerman, 2002). There is a credible motivational logic here as well. An environment that calibrates challenge to competence and leaves meaningful choices to the learner speaks directly to the needs for competence and autonomy that self-determination theory places at the root of intrinsic motivation (Deci & Ryan, 2000). The analytic dashboards typical of these systems make a learner's trajectory visible and thereby invite metacognitive ownership.

Two cautions temper this picture. First, much of the engagement and self-regulation evidence rests on self-reported measures rather than on independent performance data; increased session duration or stated enthusiasm is not, by itself, evidence of durable learning. Second, novelty effects and publication bias are real risks in a fast-moving field. The generative-AI evidence base is young, and its more dramatic claims should be held provisionally.

5. What AI Does and Does Not Do Well: The Micro-Macro Asymmetry

The most pedagogically important finding to emerge from the comparative literature is not that AI helps, but that it helps unevenly, and that the pattern of unevenness is systematic. Adaptive, AI-mediated instruction shows its largest advantages in rule-governed, practice-sensitive, micro-level dimensions of language: phonological accuracy, morphosyntactic correctness, lexical range, and the fluency that comes from high-volume, low-anxiety rehearsal. In these areas the machine's capacity to detect deviation instantly and to demand repeated re-attempts is decisive.

The picture changes at the macro level. The construction of an argument, the logical ordering of ideas across a text, the development of rhetoric, and the culturally attuned calibration of a message to an audience are competences on which current systems offer no comparable, well-evidenced advantage over capable human instruction. We state this as a reading of the evidential landscape rather than as a demonstrated null result: the outcome measures that dominate the meta-analytic corpus reviewed in Section 4 are discrete, well-specified skills, and measures of rhetorical organisation or intercultural pragmatics are rare within it. What can be said with confidence is that the case for delegating macro-level instruction to the machine has not been made. The asymmetry is also intelligible: micro-level features are well specified, abundantly represented in training data, and amenable to local correction; macro-level competence is holistic, context-dependent, and bound up with social and cultural knowledge that a model approximates only superficially.

This asymmetry carries a clear design implication, and it is the empirical hinge of the chapter's argument. An optimal architecture does not delegate the whole of instruction to the machine. It assigns to the adaptive system the routine, high-frequency work of grammatical, phonological, and lexical practice, where the evidence for AI is strongest, and reserves for the human teacher argumentation, philosophical and intercultural discussion, register and nuance, and the structural organisation of extended discourse. The division of labour follows the contours of the evidence, a principle to which Section 8 gives concrete pedagogical form.

6. The Turkish Research Ecosystem: Trends and Methodological Gaps

Situating these technologies within a specific national ecosystem shows how AI-supported adaptation is being modelled, received, and might eventually be integrated into policy. For Turkey, a small but instructive body of synthesis research now exists. Güyer and Çebi (2015) reviewed the Turkish literature on adaptive educational hypermedia (postgraduate theses, articles in Turkish educational-science journals, and SSCI-indexed articles with Turkish addresses) and mapped both the field's growth and its persistent deficiencies. Gültekin and Burak (2019), analysing the seventeen master's

and doctoral theses registered with the National Thesis Centre between 2012 and 2018, found a corpus devoted overwhelmingly to testing the effect of an adaptive learning environment on selected learner outcomes.

The methodological profile of this corpus is strikingly uniform. Most theses adopt quantitative, experimental or quasi-experimental designs; data are gathered mainly through scales and achievement tests, with qualitative techniques such as interview and observation used only rarely; and samples are drawn predominantly from higher education (Gültekin & Burak, 2019). In their modelling, Turkish adaptive designs have tended to rely on rule-based learner modelling with prior-knowledge level as the dominant individual-difference criterion (Gültekin & Burak, 2019), a choice that, in light of Section 3.1, turns out to be the empirically defensible one.

The point matters because much of the wider adaptive-learning literature took a different route: content analyses of the 2005–2014 period show learning-style-based personalisation to have been among the field's most popular design paradigms (Özyurt & Özyurt, 2015). Where gains are attributed to matching content to visual or verbal learning styles, the meshing hypothesis they presuppose lacks credible empirical support; the more defensible interpretation is that richer multimodal materials, prior-knowledge fit, or simple novelty drove the effect. Reporting such results uncritically risks propagating a neuromyth.

The synthesis studies converge on three substantive gaps. The first is a near-absence of research below higher education: Gültekin and Burak (2019) found too few studies at primary and secondary levels to support any generalisation, which leaves any move to integrate adaptive systems into national basic-education policy without a sound empirical basis. This is a serious matter, given how markedly children's and adolescents' acquisition processes, cognitive flexibility, and socio-psychological relationship with technology differ from those of adult university students. The second is a deficit of qualitative depth: studies focused on numerical outcomes alone cannot illuminate the texture of the learning process or the learner's sense of the human-machine relationship, and both reviews call for mixed-methods and qualitative designs (Güyer & Çebi, 2015; Gültekin & Burak, 2019). The third gap is one we note as an observation from our own vantage point in language education rather than as a finding of the reviews: the

Turkish adaptive-learning literature concentrates overwhelmingly on English, leaving the teaching of French and other languages almost entirely unstudied, together with the mediation-centred competences foregrounded by the CEFR Companion Volume (Council of Europe, 2020). The omission is consequential, because the CEFR's conception of mediation (facilitating access to meaning, bridging linguistic and cultural distance, and supporting the co-construction of understanding) demands precisely the kind of integration of human pedagogical judgement that goes beyond the micro-level adaptation typical of earlier systems.

7. Ethics, Bias, Privacy, and Epistemic Justice

Whatever their pedagogical promise, adaptive systems run on a fuel that is anything but neutral: granular data about real learners. To model a learner's pace, gaps, and processing, an ALMS must continuously collect, process, and store when the learner connects, how long they hesitate on each item, the typology of their errors, their reading interests, and, in speaking modules, the prosodic traces captured through the microphone. This continuous, micro-level extraction raises serious and well-founded concerns about privacy, third-party sharing, and security. Compliance with data-protection regimes (in Europe the General Data Protection Regulation; in Turkey the Law on the Protection of Personal Data, KVKK), together with anonymisation, consent-based governance, and encryption, is a legal and ethical obligation rather than an optional refinement.

Of the ethical challenges, algorithmic bias is the most charged, and it is here that the rhetoric of technological neutrality does its greatest damage. The evidence is concrete. Analysing thousands of matched utterances, Koenecke and colleagues (2020) found that five leading commercial ASR systems produced an average word error rate of 0.35 for Black speakers against 0.19 for white speakers, a near-doubling traceable to the underlying acoustic models. Their study concerned first-language speakers of African American English in the United States, so its extension to language learners is, strictly speaking, an inference. It is, however, a well-motivated one: the mechanism their analysis exposes, namely acoustic and language models trained overwhelmingly on dominant varieties, applies wherever a learner speaks with a regional, minoritised, or L2 accent, and it predicts that such

learners run a heightened risk of having their output labelled erroneous or unintelligible. Establishing the size of that risk for L2 speakers is itself a research task the field has yet to complete.

Read through Bourdieu's (1991) sociology, this is not a technical glitch but the automation of what he called the legitimate language: the variety endowed with symbolic and linguistic capital on a particular linguistic market. An ASR system that rewards the prestige accent and penalises the stigmatised one does not merely err; it encodes and naturalises an existing hierarchy of linguistic capital, conferring the appearance of objective, technical authority on what is in fact a social judgement. Read through Spivak's (1988) postcolonial critique, the same mechanism enacts a form of epistemic violence: the speaker whose variety the machine cannot hear is, in effect, denied the standing of a legitimate speaking subject. The question that animates the bias literature is, in this register, a reformulation of Spivak's own: can the subaltern speaker be heard by the machine, and on whose terms?

Framed thus, fairness-aware datasets and the representation of linguistic diversity are not engineering niceties but conditions of educational justice. Without such attention, AI in education risks deepening rather than narrowing the digital divide. The black-box problem compounds the difficulty: the models that route learners are mathematically opaque to the point that neither learner nor teacher can reconstruct why a level was lowered or a text scored poorly. The only defensible response is to require explainable AI: transparency frameworks that report, in human-interpretable terms, the grounds of consequential decisions. Finally, the capacity of generative models to produce entire assignments autonomously has destabilised conventional notions of academic integrity. The implication is not an arms race of detectors but a reorientation of assessment from product to process, so that evaluation shifts from the submitted artefact to the learner's demonstrated capacity to reason about, compile, and defend a text orally.

8. The Mediating Teacher and a Hybrid Pedagogy

The recurrent error of techno-enthusiasm is to suppose that AI can assume the whole of the teacher's function and that classrooms can be, in effect, automated. Language learning, however, is not the mechanical encoding of

syntactic rules and lexical items into memory. Language is the bearer of culture, a site of identity construction, and a process of social negotiation. The evidence reviewed above shows AI to be a tireless and often excellent assistant in diagnosing weaknesses, correcting grammatical matrices, and sharpening discrete skills. It shows, equally, that the machine cannot meet the gaze of an anxious learner with the warmth that moment requires, cannot recount a cultural anecdote with the texture of lived experience, and cannot offer the empathetic resonance that the expression of difficult personal meaning in a new language sometimes demands. The teacher's work of building socio-cultural context, transmitting value, and providing socio-emotional support is, on present evidence, not replicable.

The constructive proposal of this chapter is therefore neither automation nor refusal but a hybrid pedagogy, best articulated through the conception of mediation advanced by the CEFR Companion Volume (Council of Europe, 2020; the mediation descriptors first appeared in the 2018 provisional edition). The CEFR's mediator is an agent who facilitates access to meaning, bridges linguistic and cultural distance, and supports the co-construction of understanding, mediating texts, concepts, and communication. In a hybrid setting the teacher mediates not only between learner and target language but between learner and machine: interpreting the system's analytics through a pedagogical and ethical filter, contextualising its feedback, and converting its diagnostic output into socio-culturally meaningful learning. This is a more demanding role than transmission, not a lesser one.

Two further commitments give this mediating role concrete pedagogical content. The first is critical AI literacy: learners should be taught not merely to use generative systems but to interrogate them, recognising their biases, testing their outputs against evidence, and grasping the linguistic-market logic that those outputs silently encode (Kasneci et al., 2023). On this view the integrity threat identified earlier becomes a pedagogical opportunity: a learner who can critique and defend an AI-assisted text has thereby exercised precisely the higher-order competence the machine cannot supply. The second is the reorientation of assessment from product toward process: weighting the trajectory of composition and the learner's capacity to reason about, revise, and orally defend a text, alongside the finished artefact, renders autonomous generation insufficient on its own and returns the locus of

evaluation to the learner's demonstrable understanding. Both commitments keep the human firmly in the loop, not as a supervisor of the machine's verdicts but as the agent who decides what is worth learning and how learning is to be judged.

What might this division of labour look like when it is built into a platform rather than bolted onto one? We close the section with a design case we know from the inside and declare as such (see the declaration of interest below): LinguaFour (n.d.), an ALMS developed in Turkey with which both authors are professionally associated. The platform was designed in line with the principles defended in this chapter, not gathered here as evidence for them, and we make no empirical performance claims on its behalf. Its architecture assigns to the system the four-skill, CEFR-aligned assessment and practice of micro-level competences (accuracy in reading, listening, and writing; transcription-based analysis of pronunciation) and reserves macro-level mediation for the teacher, to whom it returns transparent, actionable analytics rather than opaque verdicts. Whether the design achieves its aims is an empirical question and, by the standards this chapter has applied to every other system, one that must be answered through independent evaluation.

9. Conclusion and Future Directions

AI-supported platforms, and adaptive learning management systems in particular, have unsettled the one-size-fits-all premise that governed language instruction for centuries. The evidence reviewed here shows that, for a constrained but important set of language-learning objectives (accuracy, pronunciation, and the fluency built through unlimited low-anxiety practice), AI and machine learning offer genuine advantages over undifferentiated instruction. The same evidence, read critically, refuses the inference that AI is a universal remedy. In macro-level competences, in the development of argumentation, in the negotiation of intercultural meaning, and in the forms of understanding that require the recognition of the learner as a social and cultural subject, the human teacher's interpretive and ethical authority remains both necessary and irreplaceable.

A defensible research and policy agenda follows directly. It should extend empirical work to the K-12 level, where the Turkish thesis corpus in

particular remains thin; privilege qualitative and mixed-methods designs that illuminate the texture of the learning process and the learner's lived experience of human-machine interaction; and attend, with greater urgency, to the teaching of languages beyond English and to the CEFR-aligned competences of mediation that the adaptive-learning literature has so far neglected. It should insist on explainable AI and on fairness-aware training with linguistically and socially diverse datasets, as conditions of educational justice rather than technical refinements. And it should position the teacher, grounded in the CEFR's conception of mediation, as the critical, interpretive, and ethically accountable agent who decides not whether to use AI in language education but how, when, and to what end.

The hybrid pedagogy defended here is neither a compromise nor a halfway house. It is, instead, a finding: the division of labour between human and machine follows the evidence about what each does well. For the applied linguist, the language teacher, and the designer of language-education systems, the work is to sustain that finding against both utopian and dystopian pressures, to keep the human in the loop, and to insist that the machine, however sophisticated its algorithms, serves learning, and not the reverse.

Declaration of Interest

3 authors are partners in AYG-GLOBAL Eğitim Teknolojileri ve Yazılım A.Ş., the company that develops the LinguaFour platform discussed as a design case in Section 8. LinguaFour is presented in this chapter solely as an illustration of design principles; no claims about its effectiveness are made or implied, and the platform played no role in the selection or interpretation of the evidence reviewed.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

During the preparation of this chapter, the authors used generative AI-assisted tools to support the literature search, the verification of bibliographic records against publisher and citation databases, and the linguistic revision of the manuscript. These tools were not used to generate the chapter's argument, analysis, or conclusions. The conceptual framework, the critical reading of the evidence, and the originality of the ideas presented are entirely the authors' own. After

using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

Bourdieu, P. (1991). *Language and symbolic power* (G. Raymond & M. Adamson, Trans.). Harvard University Press. (Original work published 1982)

Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume*. Council of Europe Publishing. https://www.coe.int/lang-cefr

Deci, E. L., & Ryan, R. M. (2000). The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry, 11*(4), 227–268.

Gültekin, M., & Burak, D. (2019). Türkiye'de uyarlanabilir öğrenme yaklaşımı kapsamında yapılan tezler üzerine bir içerik analizi [A content analysis of theses conducted within the scope of the adaptive learning approach in Turkey]. *Eğitim Teknolojisi Kuram ve Uygulama, 9*(2), 438–462.

Güyer, T., & Çebi, A. (2015). Content analysis of studies conducted on adaptive educational hypermedia environments in Turkey. *Education and Science, 40*(178), 61–83.

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, Article 102274.

Kirschner, P. A. (2017). Stop propagating the learning styles myth. *Computers & Education, 106*, 166–171.

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J. R., Jurafsky, D., & Goel, S. (2020). Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences, 117*(14), 7684–7689.

Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: A meta-analytic review. *Review of Educational Research, 86*(1), 42–78.

Li, M., Wang, Y., & Yang, X. (2025). Can generative AI chatbots promote second language acquisition? A meta-analysis. *Journal of Computer Assisted Learning, 41*(4), Article e70060.

LinguaFour. (n.d.). *LinguaFour: AI-supported adaptive language assessment and learning platform* [Web platform]. AYG-GLOBAL Eğitim Teknolojileri ve Yazılım A.Ş. https://www.linguafour.com

Ma, W., Adesope, O. O., Nesbit, J. C., & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of Educational Psychology, 106*(4), 901–918. https://doi.org/10.1037/a0037123

Mettler, E., Massey, C. M., & Kellman, P. J. (2011). Improving adaptive learning technology through the use of response times. In L. Carlson, C. Hölscher, & T. F. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 2532–2537). Cognitive Science Society.

Özyurt, Ö., & Özyurt, H. (2015). Learning style based individualized adaptive e-learning environments: Content analysis of the articles published from 2005 to 2014. *Computers in Human Behavior, 52*, 349–358.

Pashler, H., McDaniel, M., Rohrer, D., & Bjork, R. (2008). Learning styles: Concepts and evidence. *Psychological Science in the Public Interest, 9*(3), 105–119.

Riener, C., & Willingham, D. (2010). The myth of learning styles. *Change: The Magazine of Higher Learning, 42*(5), 32–35.

Spivak, G. C. (1988). Can the subaltern speak? In C. Nelson & L. Grossberg (Eds.), *Marxism and the interpretation of culture* (pp. 271–313). University of Illinois Press.

Steenbergen-Hu, S., & Cooper, H. (2014). A meta-analysis of the effectiveness of intelligent tutoring systems on college students' academic learning. *Journal of Educational Psychology, 106*(2), 331–347.

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist, 46*(4), 197–221.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory Into Practice, 41*(2), 64–70.

Agentic AI in Language Education: Designing Adaptive, Personalised, and Interactive Learning Experiences

Tuncer Can[1] & *İrfan Şimşek*[2]

1. Introduction: From Tools That Answer to Systems That Act

For most of the history of computer-assisted language learning (CALL), the computer has been a responsive object. It waited for a click, matched a string, delivered a score. Generative artificial intelligence (GenAI) changed the quality of the response but not its logic: the learner prompts, the system produces, the exchange ends. What is now described as *agentic AI* changes the logic itself. Agentic systems are built to pursue goals across time rather than to answer questions one at a time; they plan a sequence of actions, call external tools, retain memory of prior states, evaluate their own output, and initiate the next step without waiting to be asked (Sibai et al., 2026). In healthcare, where adoption has been rapid and closely scrutinised, the shift has been characterised precisely as a move "from passive, user-initiated tools to proactive agentic AI systems that are capable of autonomous, multi-step actions" (Chu et al., 2026, p. 1).

Language education is an unusually consequential place for this shift to land. Reviews of artificial intelligence in higher education have repeatedly found that language learning is the single most common subject domain in which AI is applied, with writing, reading, and vocabulary acquisition

[1] Assoc.Dr., Istanbul University Cerrahpaşa, Hasan Âli Yücel Faculty of Education, Department of English Language Teaching, ORCID: 0000-0001-8145-0772, *tcan@iuc.edu.tr*, Türkiye

[2] Prof. Dr., Istanbul University Cerrahpaşa, Hasan Âli Yücel Faculty of Education, Department of Computer Education and Instructional Technologies, ORCID: 0000-0002-7481-5830, *irfan@iuc.edu.tr*, Türkiye

dominating the applications (Crompton & Burke, 2023; Zawacki-Richter et al., 2019). The reason is structural rather than accidental: language learning requires enormous quantities of individualised interaction, immediate and contingent feedback, and repeated production under low-stakes conditions — exactly the resources that are scarcest in ordinary classrooms and most plausibly scalable through automation. A synthesis of research on AI-enabled adaptive learning systems from 2010 to 2022 found consistently positive effects on learner outcomes across 45 independent studies (X. Wang et al., 2024), and a meta-analysis of 41 experimental and quasi-experimental studies of GenAI chatbots in second language acquisition (SLA) reported a moderate-to-large pooled effect (ES = 0.576, 95% CI [0.385, 0.768], $p < .001$) across 3,515 participants (Li et al., 2025).

Yet the same period has produced a serious counter-literature. A randomised study of undergraduate essay writing found that learners supported by ChatGPT produced better texts but engaged in what the authors called *metacognitive laziness* — offloading evaluative and regulatory work to the system rather than performing it themselves (Fan et al., 2024). Reviewing this and related evidence, Yan et al. (2025) argue that education must now distinguish carefully between *performance gains* and *learning*, because generative tools reliably deliver the former without necessarily producing the latter. An experimental study of reading and writing cycles found accuracy reductions of 25.1% under complete reliance on AI for writing and 12% under AI-assisted reading (Ju, 2023).

This chapter argues that the resolution of that tension is a design problem, not a technology problem, and that agentic AI makes the design problem both more urgent and more tractable. More urgent, because an autonomous system that plans and acts can offload learner cognition far more comprehensively than a chatbot that waits to be prompted. More tractable, because an agentic architecture — unlike a single prompt — can be given an explicit pedagogical policy, a learner model, a memory of what a learner has already mastered, and guardrails that decide when *not* to help.

The chapter moves from theory to practice in five stages. Section 2 defines agentic AI and situates it on a continuum of learner–system agency. Section 3 grounds the discussion in the theoretical traditions that language educators already use: sociocultural theory, interactionist SLA, motivational

psychology, self-regulated learning, and sociomaterial accounts of distributed agency. Section 4 synthesises the empirical evidence, including its contradictions. Section 5 converts that evidence into seven design principles, and Section 6 into a reference architecture. Section 7 addresses classroom implementation, teacher competence, assessment, and ethics. Section 8 presents ApiLingua as a case study in which the principles are traced through an operating platform.

2. What "Agentic" Means, and Why It Is Not Just a Better Chatbot

2.1 The anatomy of an agent

Agentic systems are usually described in terms of five capabilities that distinguish them from single-turn generative models: planning (decomposing a goal into ordered sub-tasks), memory (persisting state across sessions), tool use (calling external systems such as speech recognisers, databases, or scoring engines), iterative reasoning (evaluating and revising intermediate output), and autonomous action (executing steps without a prompt for each) (Sibai et al., 2026). Where a generative model produces one-shot text, agentic AI "operates through iterative perception" and adaptive behaviour directed at a goal (Sibai et al., 2026, p. 4).

In education this profile is being realised through multi-agent architectures in which specialised agents divide the pedagogical labour. Recent systems illustrate the range: for instance, GraphMASAL uses a knowledge graph to coordinate agents that adapt to learners' "complex knowledge states (multi-source) and diverse goals (multi-sink)" (Zeng et al., 2025, p. 1); multi-agent learning-path planners generate transparent and explainable route recommendations through a curriculum (Xu et al., 2026); and the Evidence–Decision–Feedback framework explicitly targets the failure mode that multi-agent tutors "operate on a 'one-size-fits-all' basis, limiting their ability to provide personalized support" (Cohn et al., 2026, p. 1). These are not incremental interface changes. They relocate instructional decision-making — what to practise next, when to intervene, how much to reveal — inside the system. Bibliometric analysis of intelligent agent-supported language learning from 2005 to 2025 documents how rapidly this trajectory has

accelerated, and how far it now extends beyond conversation into personalised, adaptive, and emotionally responsive instruction (M. Liu, 2026).

2.2 A continuum of agency

It is more useful to treat agency as a continuum than as a binary. Table 1 sketches five levels, each defined by who holds the initiative for the core instructional decisions.

Table 1. A continuum of system agency in language learning environments

Level	System behaviour	Who decides what happens next	Typical instance
0. Reactive resource	Delivers fixed content; scores closed items	Curriculum designer	Traditional courseware, drill apps
1. Responsive generator	Produces language on demand	Learner, per prompt	ChatGPT used ad hoc
2. Configured tutor	Follows a designed dialogue or task script	Teacher/designer, in advance	Task-based chatbot, roleplay prompt
3. Adaptive system	Selects content from a learner model	System, within a fixed item pool	Adaptive courseware, spaced repetition
4. Agentic system	Plans multi-step sequences, calls tools, monitors progress, initiates action	System, within a pedagogical policy	Multi-agent tutor with planner and memory

The pedagogical stakes rise with each level. At Level 1 the learner remains the planner; the risk is unstructured, low-quality practice. At Level 4 the system becomes the planner; the risks are opacity, loss of teacher authority, and — most importantly for SLA — the substitution of system regulation for learner self-regulation. A structured synthesis of the move from generative to agentic AI in higher education names exactly these as the field's central tensions: instructional design, human agency, and governance (Nirban et al., 2026).

2.3 Why the shift matters for language educators specifically

Three features of language learning make agentic design particularly attractive and particularly hazardous.

First, *interaction is the curriculum*. Unlike domains where content transmission dominates, second language (L2) development depends on repeated communicative exchange. An agent that can sustain a coherent, contingent conversation across weeks — remembering a learner's recurring errors, their interests, their level — offers something no static courseware can.

Second, *feedback is the bottleneck*. Teachers cannot provide individualised corrective feedback on every utterance and every draft. Agentic systems can, and the evidence suggests the feedback is often usable (Hartshorn & Pack, 2026; Tang et al., 2026).

Third, *language is identity-laden*. Language use encodes social positioning, cultural affiliation, and voice. Systems trained on dominant corpora tend to produce standard forms rather than the pluralistic usage that reflects learners' sociolinguistic realities (S. Lee et al., 2025), and generative tools can steer interactions and "promote specific ways of thinking" in ways learners do not notice (Darvin, 2025, p. 11). Automating more of the instructional loop automates more of this influence.

3. Theoretical Foundations

An agentic system embodies a theory of learning whether or not its designers intend one. This section makes the candidate theories explicit, because each generates different, testable design commitments.

3.1 Sociocultural theory: scaffolding, contingency, and fading

Vygotsky's (1978) account of the zone of proximal development remains the most widely used frame in GenAI-assisted L2 research; a synthesis of 73 empirical studies of GenAI-assisted L2 writing found sociocultural, cognitive, and sociocognitive theories to be the most frequently invoked (Yang et al., 2025). Its central claim is that development occurs through assisted performance, where assistance is *contingent* — calibrated to what the

learner cannot yet do alone — and *temporary*, withdrawn as competence grows.

This has an uncomfortable implication for agentic design. A system optimised to maximise task success will provide maximal, permanent assistance; a system designed for development must provide minimal sufficient assistance and then remove it. Studies grounded in Vygotskian scaffolding have begun to test AI-mediated feedback against this standard directly: an eight-week quasi-experimental comparison of teacher feedback and AI-assisted peer feedback among 61 university students analysed both modes through a sociocultural lens, examining their differential effects on lexical diversity and syntactic complexity (Tang et al., 2026). The design question is not whether an agent can scaffold, but whether it can *fade*.

3.2 Interactionist SLA: input, negotiation, and output

Long's (1996) interaction hypothesis holds that acquisition is promoted when communication difficulty forces learners to negotiate meaning, making input comprehensible and pushing learners to modify output. A meta-synthesis of 57 studies on voice-based AI chatbots in English learning found that this framework, alongside task-based language teaching (TBLT), was the dominant theoretical grounding: researchers investigated whether chatbots "would increase interaction in L2 … by offering favorable circumstances for the negotiation of meaning, interaction, input, and output" (Koç & Savaş, 2024, p. 8).

The interactionist frame generates a sharp critique of over-accommodating agents. If the agent always understands, never signals non-comprehension, and never requires clarification, the interactional pressure that drives modification disappears. Task-based studies show what the alternative looks like: comparing ChatGPT-based and peer interaction for learning L2 speech acts among Korean EFL learners, Kim and Ku (2025) examined how task structure — not merely the presence of AI — shaped pragmatic learning, while task-based chatbots have been shown to enhance speaking proficiency and open translanguaging space (Shafiee Rad & Roohani, 2025).

3.3 Motivation and affect: autonomy, competence, relatedness, and willingness to communicate

Self-determination theory posits three innate psychological needs — competence, autonomy, and relatedness — whose satisfaction yields self-motivation and wellbeing and whose thwarting diminishes both (Ryan & Deci, 2000). Agentic systems interact with all three. They can support competence through calibrated difficulty and immediate feedback; they can support autonomy through genuine choice; but they can equally undermine autonomy by making the choices themselves, which is precisely what an autonomous planner is built to do.

The affective evidence in language education is encouraging but conditional. An intervention study with Chinese English learners found AI integration associated with improved achievement emotions and willingness to communicate (WTC) (Y. Chen et al., 2025); a systematic review of 37 empirical studies of learner emotions in AI-assisted ESL/EFL learning reported decreased foreign language anxiety alongside improvements in WTC and enjoyment in AI-engaged groups, with no significant change in controls (Zhang & Liu, 2025). A narrative meta-synthesis of Asian EFL contexts reached comparable conclusions about chatbot-assisted learning and WTC (Waluyo & Pratiwi, 2025), and a comparative study of AI chatbots against teachers and peers examined speaking performance, WTC, self-perceived communicative competence, and anxiety together (C. Wang & Zou, 2025). Importantly, the effect is scenario-dependent: interaction scenarios shape learners' technology acceptance and their willingness to communicate *with* AI, since scenarios "define the communicative event by shaping the roles, expectations, and goals assigned to the learner and the AI" (Cui et al., 2025, p. 2). Design, not technology, produces the affective benefit. Research combining self-determination theory with teacher scaffolding likewise finds that AI-driven support and human scaffolding interact rather than substitute for one another in shaping motivation and proficiency (Yi & Chen, 2025).

3.4 Self-regulated learning and the offloading problem

The most serious theoretical challenge to agentic design comes from research on self-regulated learning. Fan et al. (2024) demonstrated experimentally that

learners with ChatGPT support outperformed comparison groups on the immediate writing task while showing weaker self-regulatory engagement, a pattern they linked to cognitive offloading — the delegation of cognitive work to external tools — becoming habitual avoidance of deliberate effort. Yan et al. (2025) generalise the point: performance gains and learning are dissociable, and generative tools optimise the former.

Language-specific studies show the same displacement in fine detail. In a semester-long quasi-experimental study of 58 EFL university students, a GenAI-assisted visually guided writing approach significantly improved writing performance, planning, and evaluation strategies, but monitoring showed no significant gains, with prompt logs revealing "a displacement pattern in which learners outsourced real-time regulation to AI validation" (M.-R. A. Chen, 2026, p. 1). Survey research across three universities (N = 300) found that generative AI improved learner autonomy, writing performance, and confidence overall, yet high-frequency users showed stronger technological dependence and weaker academic integrity awareness (Miranty et al., 2026).

The design implication is counter-intuitive but clear: an agentic tutor must sometimes withhold. Metacognitive scaffolds are one mechanism — Torsani (2026) reports on a metacognitive checklist used to develop an in-service language teacher's self-regulation with GenAI, and metacognitive scripts have been proposed to mitigate "epistemic debt" in AI-scaffolded learning more generally (Sankaranarayanan, 2026).

3.5 Distributed agency and sociomateriality: who is acting?

Finally, agency itself needs theorising. Godwin-Jones (2024) argues that ecological theories such as sociomaterialism are needed to examine "the shared agency that develops through close user–AI interactions," noting that AI's statistical model of language constrains its capacity to handle sociocultural aspects of language use. Darvin (2025) extends this through a critical sociomaterial lens: agency "is not something possessed solely by humans but is distributed across sociotechnical systems," and learners must therefore recognise GenAI's power to steer interaction rather than "fully delegating the production of meaning and texts to technology" (p. 5).

This reframes the design goal. The aim is not to build an agent that acts *for* the learner, nor merely one that responds *to* the learner, but a configuration in which agency is distributed deliberately — with the learner retaining the decisions that constitute learning, and the teacher retaining the decisions that constitute teaching.

4. What the Evidence Supports, and Where It Conflicts

4.1 Aggregate effects

The quantitative picture is positive but heterogeneous, and it now rests on several independent syntheses. An early systematic review of chatbot-supported language learning established both the promise and the methodological weakness of the first generation of studies (Huang et al., 2021); a subsequent meta-analysis of chatbot-assisted language learning confirmed positive but variable performance effects (F. Wang et al., 2024); and a meta-analysis of robot-assisted language learning found that interventions targeting L2 speaking had produced contradictory results requiring synthesis to interpret (Y. Wang et al., 2025). Li et al. (2025) synthesised 41 studies (48 effect sizes, N = 3,515) and found GenAI chatbots produced a moderate-to-large effect on SLA (ES = 0.576), with moderator analysis showing no between-group differences by educational stage or learning environment but differences on other dimensions. X. Wang et al. (2024) found AI-enabled adaptive learning systems effective across 45 studies from 2010–2022. These sit within a longer tradition: VanLehn's (2011) analysis established that intelligent tutoring systems approach the effectiveness of human tutoring far more closely than early estimates suggested — a benchmark that agentic systems inherit and must be measured against rather than allowed to replace.

Crucially, none of these syntheses evaluates *agentic* systems as defined in Section 2. They evaluate reactive, responsive, and adaptive systems (Levels 0–3). The evidence base for Level 4 in language education is, at the time of writing, largely architectural and preliminary rather than experimental (Cohn et al., 2026; Nirban et al., 2026; Xu et al., 2026; Zeng et al., 2025). This is the single most important caveat for practitioners: enthusiasm for agentic design currently outruns evidence for it.

4.2 Writing and feedback

Writing is the best-evidenced skill area. A quasi-experimental study in a 15-week intensive English programme (n = 41) compared AI-based dynamic written corrective feedback with teacher-provided feedback on accuracy, fluency, complexity, and functional adequacy, also examining teacher and student sentiment (Hartshorn & Pack, 2026). Comparative work on feedback modes finds that AI, peer, and teacher feedback distribute differently across writing components (Zeevy-Solovey, 2024), and that AI-assisted peer feedback and teacher feedback affect lexical diversity and syntactic complexity differently, suggesting integrated rather than substitutive models (Tang et al., 2026). Learner-facing designs matter: using ChatGPT as a "writing buddy" for revision in advanced L2 German writing produced observable changes in the revision process (Strobl et al., 2024), and chatbot-supported revision has been shown to move learners from error correction toward lexical sophistication (Chon & Shin, 2026).

For assessment, LLM ratings show promise under constrained conditions. Sawaki et al. (2025) compared instructor ratings of 97 L2 summaries with LLM estimates on checklist items and found consistency depended systematically on prompt design — the amount of information supplied and the order in which rating and rationale were generated. This is an important practical finding: agentic assessment is not a property of the model but of the scaffolding around it.

4.3 Speaking, pronunciation, and oral interaction

Speech technology is the enabling layer for agentic oral practice. Reviews of automatic speech recognition (ASR) in L2 learning (Farrús, 2023) and of ASR in pronunciation teaching (McCrocklin & Levis, 2025) document both the growth and the limits of the technology, while empirical work finds ASR with peer correction improves L2 pronunciation and speaking (Sun, 2023). Purpose-built systems show what agentic oral practice can look like in situ: a voice-based chatbot designed for interactive, real-world speaking tasks supported fluency, motivation, and engagement among undergraduate learners (Oh et al., 2026), and task-based chatbots have improved speaking proficiency while shaping attitudes (Shafiee Rad & Roohani, 2025). ASR has also been used to personalise vocabulary learning by

adapting to individual learners rather than restricting practice to typed responses (Wilschut et al., 2022).

4.4 Sustained use, informal learning, and the limits of the app

A fourth strand matters for any platform-based deployment: whether learners keep using the system. MALL research has been criticised for an overemphasis on devices at the expense of socially situated practice (Pérez-Paredes & Zhang, 2022), and adoption varies sharply across populations, including older adults engaged in self-directed learning (Puebla et al., 2021). Continuance intention in MALL applications is driven by identifiable factors rather than by novelty alone (Faozi & Handayani, 2023). Meanwhile, research on informal digital learning of English (IDLE) shows that much consequential practice happens outside any course: learners differ systematically in the purposes they bring to informal digital engagement (Yao et al., 2026), affective factors such as L2 pride predict involvement (Zou et al., 2025), and teachers face ecological barriers — half of them at the exo-system level — when trying to integrate informal digital learning into instruction (Zadorozhnyy et al., 2025).

4.5 Points of agreement

Across this literature, four claims are broadly agreed.

i. **Access and volume.** AI systems substantially increase the quantity of individualised interaction and feedback available to learners (Crompton et al., 2024; Kohnke et al., 2023).

ii. **Affective benefit under low stakes.** Non-judgemental practice reduces anxiety and raises WTC for many learners (Y. Chen et al., 2025; Zhang & Liu, 2025).

iii. **Design dependence.** Outcomes vary more by pedagogical configuration than by model capability (Cui et al., 2025; Hsu, 2025; Sawaki et al., 2025).

iv. **Teacher competence is decisive.** Effective integration presupposes teacher AI literacy and assessment literacy (Kildė, 2024; Lu et al., 2025; Vogt, 2026).

4.6 Points of contradiction

Four genuine disagreements structure the field, and honest design must engage all of them.

Learning versus performing. Meta-analytic evidence of positive effects (Li et al., 2025) coexists with experimental evidence of reduced learning under reliance (Fan et al., 2024; Ju, 2023; Yan et al., 2025). The likeliest reconciliation is that outcome measures differ: immediate task quality improves while durable, transferable competence may not.

Autonomy: amplified or eroded? Configurational analysis of 43 undergraduates found that the *absence* of GenAI-supported language learning consistently predicted failure, with learner autonomy and GenAI support jointly producing gains (Hsu, 2025). Against this, high-frequency use correlates with technological dependence (Miranty et al., 2026), and monitoring is displaced to AI validation (M.-R. A. Chen, 2026). Autonomy appears to be amplified when the learner directs the tool and eroded when the tool directs the learner — which is precisely the variable that agentic design changes.

Revolution or evolution. Pegrum (2025) argues that the field should expect educational *evolution* rather than revolution, since technology is one element in a large ecosystem, and warns of pedagogical, social, and environmental risks. This tempers architectural enthusiasm without denying capability.

Whose English? GenAI has "pedagogical limitations in multilingual contexts, often generating standard English forms rather than reflecting the pluralistic usage that represents diverse sociolinguistic realities" (S. Lee et al., 2025, p. 85). An agent that autonomously selects models of correct usage is making a sociolinguistic decision at scale.

Table 2. *Contested claims and their design consequences*

Contested claim	Supporting evidence	Contrasting evidence	Design consequence
AI improves L2 outcomes	Li et al. (2025); F. Wang et al. (2024); X. Wang et al. (2024)	Fan et al. (2024); Ju (2023)	Measure delayed, unaided performance
AI supports autonomy	Hsu (2025)	M.-R. A. Chen (2026); Miranty et al. (2026)	Keep planning decisions with the learner
AI feedback rivals teacher feedback	Hartshorn & Pack (2026); Sawaki et al. (2025)	Tang et al. (2026)	Integrate modes; do not substitute
AI provides authentic models	Kohnke et al. (2023)	Darvin (2025); S. Lee et al. (2025)	Curate varieties; teach critique

5. Seven Design Principles for Agentic Language Learning

The theory and evidence above converge on a small set of design commitments. Each principle below states a rule, names its theoretical warrant, specifies the agent behaviour it requires, and identifies the failure mode it prevents.

P1. Pedagogical policy before autonomy. An agent's autonomy must be bounded by an explicit, inspectable pedagogical policy that states what the system may decide, what it must escalate, and what it must never do. This is the educational analogue of the risk-stratification frameworks proposed for agentic AI in high-stakes professional domains, where autonomy is calibrated to risk and task reversibility (Chu et al., 2026). *Failure mode:* an optimiser that maximises engagement or task completion rather than acquisition.

P2. Contingent scaffolding with designed fading. Support must be calibrated to the learner's current independent capability and systematically withdrawn as competence grows (Vygotsky, 1978). Operationally, the agent should hold a fading schedule per objective — full model, partial hint, prompt to self-correct, no support — and advance it on evidence of unaided success. The same logic governs when material returns: distributed retrieval practice across a semester improves L2 vocabulary learning more

than massed review (Nakata et al., 2020), an effect that depends on retrieval conditions and learner proficiency (Maie et al., 2025) and that combines with related memory manipulations in multidomain designs (Karataş et al., 2021; Rice & Tokowicz, 2019). *Failure mode:* permanent maximal help, which produces performance without learning (Yan et al., 2025).

P3. Productive difficulty in interaction. Because negotiation of meaning drives acquisition (Long, 1996), the agent should sometimes signal partial non-comprehension, request clarification, and push learners to modify output rather than silently repairing every utterance. Task design, not conversational fluency, is the active ingredient (Kim & Ku, 2025; Koç & Savaş, 2024). *Failure mode:* frictionless conversation that generates comfort without pressure to restructure interlanguage.

P4. Learner-held planning decisions. The agent may propose a path; the learner should authorise it. Explainable path planning makes this feasible: multi-agent planners can generate learner-facing, transparent route recommendations rather than opaque sequencing (Xu et al., 2026), and explainable learner models allow learners to inspect and correct the system's beliefs about them (J. Lee, 2021). *Failure mode:* erosion of autonomy and the displacement of monitoring to AI validation (M.-R. A. Chen, 2026).

P5. Metacognitive instrumentation. Every agentic interaction should generate a metacognitive residue: a prediction before the task, a self-assessment after it, a comparison with the system's judgement. Checklists and scripts of this kind are being tested precisely to counter offloading (Sankaranarayanan, 2026; Torsani, 2026), and prompt literacy — the capabilities needed for iterative, deliberate interaction with GenAI — is itself teachable and scaffoldable in language learning contexts (Tour & Zadorozhnyy, 2025). *Failure mode:* metacognitive laziness (Fan et al., 2024).

P6. Teacher-in-the-loop authority. Teachers must retain the ability to see, override, and redirect agent decisions. This is not a courtesy; teacher agency is dynamic and context-bound, shaped across iterational, practical-evaluative, and projective dimensions (Weng & Fu, 2025), and it is teachers who function as the primary regulators of AI use where institutional policy is absent (Avsheniuk et al., 2026). *Failure mode:* deprofessionalisation and the drift of curricular authority into vendor defaults.

P7. Assessment integrity through process evidence. When production can be delegated, product-only assessment loses validity. Teachers report that convergence of student output toward a single AI-modified form threatens task authenticity, performance authenticity, and interpretability (Avsheniuk et al., 2026), while holistic and interactive assessment components are proposed as mitigation (Cong-Lem & Nguyen, 2024). Agentic systems can supply what static assessment cannot: process logs, revision histories, and interaction traces. *Failure mode:* an integrity arms race built on unreliable detection.

6. A Reference Architecture

Principles become practice through architecture. The following layered design is offered as a template that institutions can adapt; it is deliberately expressed in pedagogical rather than vendor-specific terms.

Layer 1 — Learner model. A persistent, inspectable representation of proficiency by skill and objective, recurring error patterns, task history, interests, and affective indicators. Two design decisions matter more than the modelling technique. First, the model should be *editable* by the learner, as demonstrated in text recommendation for language learning, where allowing learners to correct the system's assumptions improved fit (J. Lee, 2021). Second, it should record evidence of *unaided* performance separately from assisted performance, since only the former licenses fading.

Layer 2 — Pedagogical policy engine. The explicit rule set implementing P1–P3: which objective is due, what support level applies, when to escalate to a human. Approaches such as Evidence–Decision–Feedback formalise this as theory-driven adaptive scaffolding rather than undifferentiated helpfulness (Cohn et al., 2026).

Layer 3 — Agent ensemble. Rather than one general assistant, a small set of specialised roles with distinct instructions and permissions (Table 3). Multi-agent designs improve response quality through specialisation, though at a real cost: concurrent agent calls compound latency through a parallel-phase maximum effect that single-agent systems do not face (Elhaimeur & Chrisochoides, 2026).

Table 3. Agent roles, SLA functions, and characteristic risks

Role	Function in the learning cycle	SLA warrant	Characteristic risk
Interlocutor	Sustains task-based interaction; negotiates meaning	Long (1996); Koç & Savaş (2024)	Over-accommodation; loss of interactional pressure
Feedback agent	Provides contingent corrective feedback on output	Hartshorn & Pack (2026)	Over-correction; feedback beyond the learner's ZPD
Curator	Selects and levels texts and tasks; schedules recycling	J. Lee (2021); Nakata et al. (2020)	Narrow, homogenised input
Assessor	Produces criterion-referenced judgements with rationale	Sawaki et al. (2025)	Prompt-dependent inconsistency
Metacognitive coach	Elicits prediction, reflection, and self-evaluation	Fan et al. (2024); Torsani (2026)	Ritualised, unread reflection prompts
Teacher liaison	Surfaces decisions, exceptions, and evidence to the teacher	Weng & Fu (2025)	Dashboard overload

Layer 4 — Tool layer. ASR and pronunciation scoring (Farrús, 2023; McCrocklin & Levis, 2025), spaced retrieval scheduling (Maie et al., 2025; Nakata et al., 2020), corpus and dictionary access, and content generation. Tool use is what separates an agent from a conversationalist: it lets the system act on the world of the learner's materials rather than only talk about them.

Layer 5 — Evaluation and guardrails. Continuous monitoring for hallucination, inappropriate difficulty, bias in the varieties of language modelled, and — distinctively for multi-agent systems — adversarial vulnerability, since security research shows that multi-agent configurations introduce weaknesses that single-agent evaluation does not reveal (Kavathekar et al., 2025). Educational deployment adds a pedagogical guardrail: an agent should refuse to complete assessed work on a learner's behalf.

Layer 6 — Teacher and institution interface. Class-level views of objective mastery, flagged learners, and agent decisions, plus policy controls. Dashboard research warns that this layer is harder than it looks: student-facing dashboards grounded in self-regulated learning theory are interpreted very differently by different learners (De Barba et al., 2022), and

dashboards designed explicitly to support metacognition require theory-driven, explainable design rather than descriptive charts (L. Chen et al., 2021). Explainability is not decoration here; it is the mechanism by which teacher authority (P6) is exercised.

7. From Architecture to Classroom

7.1 A lesson-level workflow

Agentic capability is best introduced inside familiar pedagogical structures rather than alongside them. A task cycle offers a natural fit.

Pre-task. The curator agent assembles level-appropriate input and pre-teaches the lexis the task will demand; the metacognitive coach elicits the learner's prediction of difficulty and a goal for the cycle.

During task. The learner performs the task with the interlocutor agent, which maintains the communicative scenario. Support is set by the policy engine at the level the learner model licenses. Where the task is oral, ASR feeds pronunciation-focused feedback; where written, the feedback agent withholds correction until the draft is complete, preserving fluency.

Post-task. The feedback agent provides focused, selective correction on a small number of targeted forms. The learner self-assesses before seeing the assessor's judgement, then compares. The teacher liaison surfaces the class-level pattern: which objectives are unstable, which learners need human intervention.

Recycling. The curator schedules retrieval of the targeted forms in later tasks, and the learner model registers unaided success or failure — the evidence that drives fading (P2).

7.2 Teacher competence

None of this is teacher-proof, and it should not be. Chapelle et al. (2024) examined materials for teaching technological pedagogical content knowledge in a professional development course introducing international English language teachers to ChatGPT, noting that open GenAI expands the TPACK teachers require without yet clarifying what teacher educators should teach. Systematic review of GenAI in foreign language teacher

education confirms an emergent, uneven research base (Kildė, 2024), and critical mapping of the field argues that teachers must be prepared for an *algorithmic* educational landscape rather than merely a new tool (Karakaya et al., 2025).

AI literacy is the organising construct. A multiple case study of three teacher candidates found AI literacy developing unevenly across four domains — technological proficiency, pedagogical compatibility, professional work, and ethical use — with critical stances shaped by evolving professional identities (Lu et al., 2025). Critical AI literacy extends this to interrogating the systems themselves (Pérez-Paredes et al., 2025; Tacelosky et al., 2025). Prompt design is the operational core of this competence: prompt literacy training changes the language learning strategies students deploy (Dillon, 2024), postgraduate writers develop identifiable prompting practices that instruction can shape (Caprario et al., 2026), and prompt engineering for language teachers is best understood as an extension of pedagogical expertise rather than a technical skill (Georgiou, 2026). Teacher affect matters too: AI anxiety is a documented barrier to adoption (Parviz & Arthur, 2025), and teachers' cognition shapes how GenAI is actually used to support engagement (Zaimoğlu & Dağtaş, 2025).

7.3 Assessment redesign

Three moves follow from P7. First, shift weight from product to process, using the interaction and revision traces agentic systems already generate. Second, add interactive components — oral defence, in-class performance — as recommended by teachers responding to integrity concerns (Cong-Lem & Nguyen, 2024). Third, build teacher language assessment literacy that explicitly incorporates AI: Vogt (2026) maps affordances (automated item generation, scalable scoring, timely individualised feedback) against challenges (overreliance, integrity threats, bias, data protection, equity) and argues these competences must be embedded in initial and in-service teacher education. Detection-based responses are the weakest option; teachers themselves report reliance on detectors alongside doubts about them (Cong-Lem & Nguyen, 2024; Özer, 2024).

7.4 Ethics, equity, and governance

Ethical questions in AI-supported education are not resolved by good intentions: fairness, accountability, transparency, bias, autonomy, agency, and inclusion require explicit treatment, and most researchers in the field are not trained to address them (Holmes et al., 2022). For language education specifically, four issues recur.

Data protection. Learner models are longitudinal records of linguistic and behavioural data. Systematic review of privacy and data protection in learning analytics shows that these issues remain unresolved even as the field matures into deployment (Q. Liu & Khalil, 2023), that adequate responses are institutional and ethical rather than merely technical (Prinsloo et al., 2022), and that student privacy is a defining ethical problem for AI in education (L. Huang, 2023). Retention limits, minimisation, and transparency are preconditions for deployment, not features.

Linguistic bias and Global Englishes. Systems that model only prestige varieties narrow learners' sense of legitimate usage (S. Lee et al., 2025). Curation of varieties should be an explicit design decision.

Equity of access. Practical constraints in AI use fall hardest on less privileged populations (Godwin-Jones, 2024). Offline-tolerant and low-bandwidth designs are equity features.

Governance. Where institutional AI policy is absent, teachers become the de facto regulators, an unsustainable arrangement (Avsheniuk et al., 2026). Structured syntheses of agentic AI in higher education identify governance as one of three central tensions alongside instruction and human agency (Nirban et al., 2026).

7.5 Evaluating agentic systems

Because performance gains and learning are dissociable (Yan et al., 2025), evaluation must be designed to separate them. A defensible minimum for institutional adoption comprises: (a) delayed, unaided post-tests rather than assisted task quality alone; (b) transfer tasks in modalities the system did not train; (c) self-regulation and metacognitive measures alongside proficiency measures (M.-R. A. Chen, 2026; Fan et al., 2024); (d) affective measures — anxiety, enjoyment, WTC — since these are among the most consistent benefits (Zhang & Liu, 2025); and (e) equity analysis across proficiency bands,

since AI-assisted gains have been found to accrue disproportionately to learners with stronger prior knowledge and skills (Ju, 2023).

8. Case Study: ApiLingua

8.1 The problem the platform names

ApiLingua (https://apilingua.com) is an AI-supported English learning application developed by Girift EdTech and available for iOS and Android. Its public positioning is unusually specific for a consumer language app. The headline question — "Understand English but can't speak?" — and the accompanying promise to "turn understanding into speaking" name a phenomenon that SLA research treats as a genuine construct rather than a marketing frame: the asymmetry between receptive competence and productive performance. In interactionist terms, comprehensible input alone does not force the syntactic processing that production requires (Long, 1996). In affective terms, the gap is sustained by anxiety and low willingness to communicate, which is precisely where AI practice partners have shown their most consistent benefit (Y. Chen et al., 2025; Waluyo & Pratiwi, 2025; Zhang & Liu, 2025). A platform that defines its target outcome as speaking, and its target user as someone who already understands, is therefore aiming at the point where the evidence for AI-mediated practice is strongest.

It is also aiming at a population that language education reaches poorly. Much consequential English practice now happens informally, outside instruction, where learners differ in purpose and motivation (Yao et al., 2026; Zou et al., 2025) and where teachers face ecological barriers to integration (Zadorozhnyy et al., 2025). Self-access platforms occupy this space, with the accompanying risk that use decays once novelty passes — continuance in mobile-assisted language learning is driven by identifiable design and experience factors rather than initial enthusiasm (Faozi & Handayani, 2023; Pérez-Paredes & Zhang, 2022).

8.2 The learning architecture as it stands

The platform's public description supports a fairly precise reading of its design. Content is organised on a CEFR spine from A1 to B2, with ten curated topics per level and skill categories covering vocabulary, grammar, reading, listening

and writing, comprising more than 6,000 expert-crafted interactive activities. The learner journey is stated in four steps — discover your level, complete tasks, practise with AI, analyse progress — beginning with a placement test and gated by a completion rule: progression requires at least 80% completion of the current topics. Two conversational components carry the productive load: an AI chatbot for writing practice returning instant, level-appropriate feedback, and a voice assistant for speaking practice through realistic dialogues with instant responses. Progress is reported through AI-generated insights, personalised feedback reports, and a dashboard tracking completion for every skill and topic.

Mapped onto the reference architecture of Section 6, the platform implements parts of four layers and leaves two largely open (Table 4). On the continuum in Table 1 it sits at Levels 2–3: a designed, adaptive environment with configured conversational partners, not an agentic system that plans across sessions. That is the analytically interesting property of this case. It shows what a well-formed pre-agentic platform already achieves, and it makes visible exactly which decisions would have to be made — and which risks accepted — for the transition to Level 4.

Table 4. ApiLingua components mapped to the reference architecture

Layer	Present in the platform	Not yet visible
1. Learner model	Placement test; per-skill and per-topic completion tracking	Per-objective mastery; assisted vs unaided evidence; learner editability
2. Policy engine	Fixed CEFR sequencing; 80% completion gate	Explicit support ladder; escalation rules
3. Agent ensemble	Writing chatbot; speaking voice assistant	Curator, assessor, metacognitive coach, teacher liaison as distinct roles
4. Tool layer	Speech interaction; activity bank; feedback reports	Spaced retrieval scheduling tied to the learner model
5. Guardrails	Level-appropriate feedback	Published policy on bias, variety curation, refusal to complete assessed work
6. Teacher interface	Learner-facing dashboard	Class views, override controls, exportable process evidence

8.3 Reading the platform against the seven principles

What the design already gets right. The CEFR spine and the completion gate operationalise P2 in coarse form: progression is contingent on evidence and content is sequenced rather than browsed. The placement test instantiates the beginning of a learner model (P4, Layer 1). Separating a writing chatbot from a speaking assistant respects a real distinction in the evidence, where written corrective feedback and oral fluency support have different research bases and different failure modes (Hartshorn & Pack, 2026; Oh et al., 2026). Level-appropriate feedback is an implicit ZPD commitment. And the four-step journey — level, tasks, AI practice, analysis — has the shape of a task cycle, which is the structure that agentic components integrate into most easily.

Where the literature predicts strain. Three gaps are visible from the public description and are characteristic of the category rather than peculiar to this platform. First, completion is a proxy for mastery, not a measure of it: the fading logic in P2 requires evidence of *unaided* success on specific objectives, which a completion percentage cannot distinguish from assisted success or from repeated exposure. Second, a chatbot tuned to be level-appropriate and instantly responsive risks the over-accommodation failure mode in P3 — comfort without the interactional pressure that drives modification (Long, 1996; Koç & Savaş, 2024). Third, the analytics are learner-facing and outcome-oriented; the teacher and institution layer specified in P6, and the process evidence required by P7, are outside the consumer product's scope.

A fourth issue is specific to the CEFR spine. Framing content by CEFR level is a claim about proficiency, and the language assessment literature is explicit that alignment claims require validity evidence rather than assertion (Papageorgiou, 2019). Mapping to CEFR levels enhances score interpretation but is frequently misread as evidence of test quality or cross-test comparability (Papageorgiou, 2022). The framework's global influence makes this an easy claim to make loosely (Sahib & Stapa, 2022), while reviews of implementation show that its effects on classroom practice are mixed and demanding for teachers (Sahib & Stapa, 2021). A platform organised on can-do levels should therefore treat CEFR alignment as an

empirical commitment: what evidence links the placement decision, the topic bands, and the feedback standards to the levels they name?

8.4 A staged transition toward agency

The roadmap below is a design proposal derived from Section 5, not a description of existing functionality. It is sequenced so that each stage is independently evaluable and so that no stage grants the system more autonomy than the previous stage's evidence supports.

Stage 1 — Evidence-grade learner model. Extend the placement test into a persistent, per-objective model recording assisted and unaided performance separately, and make it learner-editable (J. Lee, 2021). This converts an 80% completion gate into a mastery decision and is the precondition for everything that follows. *Success criterion:* the model predicts unaided performance on held-out tasks better than completion rate does.

Stage 2 — Policy-governed feedback. Introduce an explicit support ladder in the writing chatbot — full model, partial hint, self-correction prompt, no support — advancing on unaided success. Prompt design should be treated as a controlled variable rather than a fixed setting, given evidence that agreement between LLM and instructor judgements depends systematically on prompt structure and the ordering of rating and rationale (Sawaki et al., 2025). *Success criterion:* accuracy gains hold when support is withdrawn.

Stage 3 — Task-based interlocutor with designed friction. Reconfigure the voice assistant from a responsive conversational partner into a task-based interlocutor with outcome-defined tasks, calibrated clarification requests, and pushed output — the configuration with the strongest SLA warrant (Kim & Ku, 2025; Koç & Savaş, 2024; Shafiee Rad & Roohani, 2025). Interaction scenarios should be varied deliberately, since scenario shapes both acceptance and willingness to communicate with AI (Cui et al., 2025), and conversational partners of this kind have a decade of design evidence to draw on (Belda-Medina & Calvo-Ferrer, 2022; W. Huang et al., 2021). *Success criterion:* increased modified output without loss of enjoyment or WTC.

Stage 4 — Metacognitive layer. Add pre-task prediction and post-task self-assessment, with explicit comparison against the system's judgement, as the structural counterweight to offloading (M.-R. A. Chen, 2026; Fan et al.,

2024; Torsani, 2026), and teach prompt literacy inside the product rather than assuming it (Tour & Zadorozhnyy, 2025). *Success criterion:* calibration between self-assessment and system judgement improves over time.

Stage 5 — Planner with learner authorisation. Only here does the system become agentic in the Level 4 sense: a planner proposes multi-session routes from the learner model, interleaving spaced retrieval of unstable objectives with new material (Maie et al., 2025; Nakata et al., 2020), but routes are shown transparently and authorised by the learner (P4), following explainable path-planning approaches (Xu et al., 2026). *Success criterion:* learners accept, and sometimes amend, proposed routes — amendment rate is the evidence that authorisation is real rather than nominal.

Stage 6 — Institutional edition. Add the teacher liaison layer: class views, override controls, and exportable process evidence for assessment (P6, P7), designed against the dashboard literature's warning that interpretation varies by learner and that explainability must be built in rather than added (L. Chen et al., 2021; De Barba et al., 2022).

8.5 Two implementation vignettes

The autonomous learner. An A2 user who understands workplace English but freezes in meetings is placed at A2, and the metacognitive coach asks what she wants to be able to do in six weeks. Her authorised route interleaves three topics with a weekly speaking task. In each task the interlocutor plays a colleague, asks for clarification when her utterance is ambiguous, and does not repair for her. Afterwards she predicts her own accuracy, sees the assessor's judgement, and watches the gap between the two narrow. Objectives she fails unaided reappear at increasing intervals.

The blended classroom. A B1 class uses the platform between lessons. The teacher sets the objectives, and the teacher liaison reports which are unstable across the class; the teacher then designs a face-to-face task around exactly those. Assessment is based on in-class performance plus the platform's process evidence — drafts, revisions, interaction logs — rather than on any product the learner could have delegated (P7).

8.6 An evaluation agenda and the limits of this case

A platform holding longitudinal, per-objective data can answer questions the field currently cannot, at a scale classroom research rarely achieves. Four studies would be of direct value beyond the product. The first is a delayed, unaided speaking post-test comparing a responsive voice assistant against a task-based interlocutor with designed friction, testing P3 against the interactionist prediction. The second is a randomised comparison of the support ladder against constant maximal feedback, with self-regulation measures alongside proficiency measures, testing whether designed fading mitigates the displacement effect documented by M.-R. A. Chen (2026). The third is a validity study of the CEFR alignment underlying placement and progression (Papageorgiou, 2019, 2022). The fourth is an equity analysis of gains across proficiency bands, given evidence that AI-assisted benefit is unevenly distributed by prior knowledge (Ju, 2023), combined with a retention analysis addressing continuance rather than initial adoption (Faozi & Handayani, 2023).

Three limits should be stated plainly. This case rests on publicly available platform information rather than on internal design documentation or independent efficacy data; no peer-reviewed evaluation of the platform exists at the time of writing, and none of the claims here should be read as evidence that it outperforms alternatives. Second, the roadmap is a normative proposal: it says what the design principles imply, not what the developer intends. Third, the longitudinal learner data that makes the evaluation agenda feasible is precisely the data that carries the governance obligations set out in Section 7.4 — a platform that can run these studies is also a platform that must be able to state its retention, minimisation, and consent commitments (L. Huang, 2023; Q. Liu & Khalil, 2023; Prinsloo et al., 2022).

9. Conclusion

Agentic AI is best understood not as a more capable tutor but as a relocation of instructional decision-making. When a system can plan, remember, act, and initiate, it acquires a share of the decisions that used to belong to teachers and learners — what to practise, how much to help, when to move on. That is why the arguments in this chapter have been about design commitments rather than model capabilities. The empirical record supports

optimism about access, affect, and feedback volume, and it supports caution about the difference between performing well and learning (Fan et al., 2024; Li et al., 2025; Yan et al., 2025). The evidence base for genuinely agentic systems in language education remains thin, which places an obligation on designers and institutions to build evaluable systems rather than impressive demonstrations.

The chapter's practical claim is that seven commitments — bounded autonomy, fading scaffolding, productive difficulty, learner-held planning, metacognitive instrumentation, teacher authority, and process-based assessment — are sufficient to convert agentic capability into pedagogical value, and that their absence is what turns capability into cognitive offloading. The ApiLingua case shows the transition is incremental: a well-structured adaptive platform already satisfies parts of the framework, and each further stage is independently testable. Agency, as the sociomaterial literature insists, is distributed across the whole configuration of learner, teacher, task, and system (Darvin, 2025; Godwin-Jones, 2024). The design question is not how much agency to give the machine, but how to arrange the whole configuration so that the learner keeps the agency that constitutes learning.

Author's note on sources. All references below were retrieved and verified through the scite.ai literature database and checked for editorial notices (retractions, corrections, expressions of concern); none of the cited records carried such a notice at the time of writing. Where an indexed record listed only the first three authors, the reference entry follows that record and uses "et al."; authors preparing this chapter for publication should expand those entries against the original articles. Where an article was indexed ahead of print, the year of the indexed record is used. Information about ApiLingua in Section 8 derives from the platform's public website (https://apilingua.com); the mapping in Table 4 and the roadmap in Section 8.4 are the author's analysis and design proposal, not descriptions of released functionality. Readers should note that the platform is a commercial product, and any author affiliation with its developer should be declared in a conflict-of-interest statement.

References

Avsheniuk, N., Seminikhyna, N., & Lutsenko, O. (2026). Homogenization and epistemic dependency: How generative AI threatens task-based language assessment in ESP contexts. *Arab World English Journal, 17*(12), 400–417.

Belda-Medina, J., & Calvo-Ferrer, J. R. (2022). Using chatbots as AI conversational partners in language learning. *Applied Sciences, 12*(17), Article 8427.

Caprario, M., Lin, L. H. F., & Li, V. W. Y. (2026). Reimagining writing instruction with GenAI: Insights into postgraduate EAP students' GenAI prompt engineering practices in Hong Kong. *Journal of English for Academic Purposes, 82*, Article 101695.

Chapelle, C. A., Baghestani, S., & Beck, J. C. (2024). Exploring open AI in language teacher professional development (pp. 217–242). Iowa State University Digital Press.

Chen, L., Lu, M., & Goda, Y. (2021). Learning analytics dashboard supporting metacognition (pp. 129–149). Springer.

Chen, M.-R. A. (2026). Generative AI, visual scaffolds, and metacognition in EFL writing. *Language Learning & Technology, 30*(1), 1–26.

Chen, Y., Zhi, Y., & Derakhshan, A. (2025). Integrating artificial intelligence (AI) into the English as a foreign language classroom: Exploring its impact on students' achievement emotions and willingness to communicate (WTC). *European Journal of Education, 60*(3).

Chon, Y. V., & Shin, D. (2026). AI chatbot as a revision aid in second language writing: From error correction to lexical sophistication. *Journal of Computer Assisted Learning, 42*(3).

Chu, B., Shen, H., & Lee, I. (2026). Agentic AI in dermatology: A call to action. *JMIR Dermatology, 9*, Article e94909.

Cohn, C., Guo, S., Rayala, S., et al. (2026). *Evidence-decision-feedback: Theory-driven adaptive scaffolding for LLM agents* [Preprint]. arXiv.

Cong-Lem, N., & Nguyen, T. T. (2024). Academic integrity in the age of generative AI: Perceptions and responses of Vietnamese EFL teachers. *Teaching English with Technology, 2024*(1).

Crompton, H., & Burke, D. (2023). Artificial intelligence in higher education: The state of the field. *International Journal of Educational Technology in Higher Education, 20*(1).

Crompton, H., Edmett, A., Ichaporia, N., et al. (2024). AI and English language teaching: Affordances and challenges. *British Journal of Educational Technology, 55*(6), 2503–2529.

Cui, Z., Yang, H., & Xu, H. (2025). The effects of interaction scenarios on EFL learners' technology acceptance and willingness to communicate with AI. *Behavioral Sciences, 15*(10), Article 1391.

Darvin, R. (2025). Identity and investment in the age of generative AI. *Annual Review of Applied Linguistics, 45*, 10–27.

De Barba, P., Oliveira, E., & Hu, X. (2022). Same graph, different data: A usability study of a student-facing dashboard based on self-regulated learning theory. *ASCILITE Publications*, Article e22168.

Dillon, T. J. (2024). Korean university students' prompt literacy training with ChatGPT: Investigating language learning strategies. *English Teaching, 79*(3), 123–157.

Elhaimeur, I., & Chrisochoides, N. (2026). *Latency and cost of multi-agent intelligent tutoring at scale* [Preprint]. arXiv.

Fan, Y., Tang, L., Le, H., et al. (2024). Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. *British Journal of Educational Technology, 56*(2), 489–530.

Faozi, F. H., & Handayani, P. W. (2023). The antecedents of mobile-assisted language learning applications continuance intention. *The Electronic Journal of e-Learning, 21*(4), 299–313.

Farrús, M. (2023). Automatic speech recognition in L2 learning: A review based on PRISMA methodology. *Languages, 8*(4), Article 242.

Georgiou, G. P. (2026). *Prompt engineering for language teachers using ChatGPT: A brief guide to techniques, applications, and effective practices* [Preprint]. MDPI.

Godwin-Jones, R. (2024). Distributed agency in second language learning and teaching through generative AI. *Language Learning & Technology, 28*(2), 5–31.

Hartshorn, K. J., & Pack, A. (2026). The effects of artificial intelligence-based dynamic written corrective feedback on second language writing and user sentiment. *RELC Journal, 57*(1), 69–88.

Holmes, W., Porayska-Pomsta, K., Holstein, K., et al. (2022). Ethics of AI in education: Towards a community-wide framework. *International Journal of Artificial Intelligence in Education, 32*(3), 504–526.

Hsu, L. (2025). A configurational exploration on the promise and perils of generative AI in project-based language learning. *International Journal of Computer-Assisted Language Learning and Teaching, 15*(1), 1–20.

Huang, L. (2023). Ethics of artificial intelligence in education: Student privacy and data protection. *Science Insights Education Frontiers, 16*(2), 2577–2587.

Huang, W., Hew, K. F., & Fryer, L. K. (2021). Chatbots for language learning—Are they really useful? A systematic review of chatbot-supported language learning. *Journal of Computer Assisted Learning, 38*(1), 237–257.

Ju, Q. (2023). *Experimental evidence on negative impact of generative AI on scientific learning outcomes* [Preprint]. Research Square.

Karakaya, K., Alpat, M. F., & Uçar, H. (2025). Preparing teachers for the algorithmic educational landscape: A critical mapping of generative AI integration in language teacher education. *Technology in Language Teaching & Learning, 7*(2), Article 102841.

Karataş, N. B., Özemir, O., Lovelett, J., et al. (2021). Improving second language vocabulary learning and retention by leveraging memory enhancement techniques: A multidomain pedagogical approach. *Language Teaching Research, 29*(1), 112–149.

Kavathekar, I., Jain, H., Rathod, A., et al. (2025). *TAMAS: Benchmarking adversarial risks in multi-agent LLM systems* [Preprint]. arXiv.

Kildė, L. (2024). The integration of generative AI in foreign language teacher education: A systematic literature review. *Pedagogika, 154*(2), 5–26.

Kim, Y., & Ku, K. W. C. (2025). Exploring generative AI as a roleplay interlocutor in L2 task-based pragmatics learning: Comparing ChatGPT-learner and technology-mediated peer interactions. *TESOL Quarterly, 59*(S1).

Koç, F., & Savaş, P. (2024). The use of artificially intelligent chatbots in English language learning: A systematic meta-synthesis study of articles published between 2010 and 2024. *ReCALL, 37*(1), 4–21.

Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for language teaching and learning. *RELC Journal, 54*(2), 537–550.

Lee, J. (2021). An editable learner model for text recommendation for language learning. *ReCALL, 34*(1), 51–65.

Lee, S., Jeon, J., & McKinley, J. (2025). Generative AI and English language teaching: A global Englishes perspective. *Annual Review of Applied Linguistics, 45*, 85–108.

Li, M., Wang, Y., & Yang, X. (2025). Can generative AI chatbots promote second language acquisition? A meta-analysis. *Journal of Computer Assisted Learning, 41*(4).

Liu, M. (2026). The rise of digital tutors: A bibliometric exploration of intelligent agent-supported language learning from 2005 to 2025. *European Journal of Education, 61*(1).

Liu, Q., & Khalil, M. (2023). Understanding privacy and data protection issues in learning analytics using a systematic review. *British Journal of Educational Technology, 54*(6), 1715–1747.

Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 413–468). Academic Press.

Lu, X., Zaitseva, A., & Troyan, F. J. (2025). AI literacy: A core practice in world language education. *Foreign Language Annals, 59*(1), 9–33.

Maie, R., Oikawa, G., & Chen, Y. (2025). Cumulative testing for L2 vocabulary learning: The impact of retrieval practice and proficiency. *TESOL Quarterly, 59*(4), 1979–2008.

McCrocklin, S., & Levis, J. M. (2025). Automatic speech recognition and pronunciation learning. *Language Teaching, 59*(3), 293–309.

Miranty, D., Ali, H., & Huda, N. (2026). Democratizing language proficiency: The impact of generative AI on learner autonomy and academic integrity in second language acquisition. *Lingeduca: Journal of Language and Education Studies, 5*(1), 1–18.

Nakata, T., Tada, S., & McLean, S. (2020). Effects of distributed retrieval practice over a semester: Cumulative tests as a way to facilitate second language vocabulary learning. *TESOL Quarterly, 55*(1), 248–270.

Nirban, H. S., Bhatia, R., & Mangal, B. (2026). *From generative AI to agentic AI in higher education: A structured literature synthesis of instructional, human agency, and governance tensions* [Preprint]. Research Square.

Oh, E. Y., Choi, J., & Song, D. (2026). Designing and evaluating a voice-based chatbot system for second language speaking fluency and learner engagement. *Journal of Educational Technology Systems, 54*(4), 761–785.

Özer, Ö. (2024). AI language models: A breach of academic integrity in online language learning? *Studies in Language Assessment*, 237–262.

Papageorgiou, S. (2019). Issues in aligning assessments to the Common European Framework of Reference. *Language Value, 6*(1).

Papageorgiou, S. (2022). Still deluded by artifices? The role of the Common European Framework of Reference in facilitating test score interpretation. *Language Teaching Research Quarterly, 29*, 57–64.

Parviz, M., & Arthur, F. (2025). AI anxiety in English language education. *International Journal of Computer-Assisted Language Learning and Teaching, 15*(1), 1–21.

Pegrum, M. (2025). From revolution to evolution: What generative AI really means for language learning. *Language Teaching, 59*(2), 180–196.

Pérez-Paredes, P., Curry, N., & Ordoñana Guillamón, C. (2025). Critical AI literacy for applied linguistics and language education students. *Journal of China Computer-Assisted Language Learning, 5*(2), 175–214.

Pérez-Paredes, P., & Zhang, D. (2022). Mobile assisted language learning: Scope, praxis and theory. *Porta Linguarum*, 11–25.

Prinsloo, P., Slade, S., & Khalil, M. (2022). The answer is (not only) technological: Considering student data privacy in learning analytics. *British Journal of Educational Technology, 53*(4), 876–893.

Puebla, C., Fievet, T., Tsopanidi, M., et al. (2021). Mobile-assisted language learning in older adults: Chances and challenges. *ReCALL, 34*(2), 169–184.

Rice, C. A., & Tokowicz, N. (2019). A review of laboratory studies of adult second language vocabulary training. *Studies in Second Language Acquisition, 42*(2), 439–470.

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist, 55*(1), 68–78.

Sahib, F. H., & Stapa, M. (2021). The impact of implementing the Common European Framework of Reference on language education: A critical review. *International Journal of Academic Research in Business and Social Sciences, 11*(11).

Sahib, F. H., & Stapa, M. (2022). Global trends of the Common European Framework of Reference: A bibliometric analysis. *Review of Education, 10*(1).

Sankaranarayanan, S. (2026). *Mitigating "epistemic debt" in generative AI-scaffolded novice programming using metacognitive scripts* [Preprint]. arXiv.

Sawaki, Y., Ishii, Y., & Yamada, H. (2025). Examining the consistency of instructor versus large language model ratings on summary content: Toward checklist-based feedback provision with second language writers. *Language Testing, 42*(4), 447–475.

Shafiee Rad, H., & Roohani, A. (2025). AI language alchemists: Unleashing task-based chatbots to enhance speaking proficiency, shape attitudes, and foster a translanguaging space. *Journal of Educational Computing Research, 63*(7–8), 1659–1688.

Sibai, N. S., Ahmed, Y., Sibaee, S., et al. (2026). *The path ahead for agentic AI: Challenges and opportunities* [Preprint]. arXiv.

Strobl, C., Menke-Bazhutkina, I., & Abel, N. (2024). Adopting ChatGPT as a writing buddy in the advanced L2 writing class. *Technology in Language Teaching & Learning, 6*(1), Article 1168.

Sun, W. (2023). The impact of automatic speech recognition technology on second language pronunciation and speaking skills of EFL learners: A mixed methods investigation. *Frontiers in Psychology, 14*, Article 1210187.

Tacelosky, K., Kasun, G. S., & Shapiro, B. R. (2025). Exploring critical AI literacy in language education: A case study. *Foreign Language Annals, 58*(4), 966–994.

Tang, J., Li, P., & Luo, R. (2026). Teacher feedback vs AI-assisted peer feedback in L2 writing: A quasi-experimental study in a Chinese university. *PLOS ONE, 21*(6), Article e0345976.

Torsani, S. (2026). Developing self-regulation for generative AI through a metacognitive checklist: A case study of an in-service language teacher. *Australian Journal of Applied Linguistics, 8*(4), Article 103284.

Tour, E., & Zadorozhnyy, A. (2025). Conceptualizing and operationalizing prompt literacy for English language learners. *Journal of Adolescent & Adult Literacy, 69*(3).

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist, 46*(4), 197–221.

Vogt, K. (2026). Generative AI technology and teacher language assessment literacy – Can language teachers keep up? *Language Testing in Focus, 13*, 19–31.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Waluyo, B., & Pratiwi, D. I. (2025). AI chatbot-assisted English learning and willingness to communicate: A narrative meta-synthesis of evidence from Asian English as a foreign language contexts. *The JALT CALL Journal, 21*(3), Article 102884.

Wang, C., & Zou, B. (2025). AI versus teachers and peers: A comparative study on speaking performance, willingness to communicate, self-

perceived communicative competence and foreign language anxiety. *IRAL – International Review of Applied Linguistics in Language Teaching*.

Wang, F., Cheung, A., & Neitzel, A. (2024). Does chatting with chatbots improve language learning performance? A meta-analysis of chatbot-assisted language learning. *Review of Educational Research, 95*(4), 623–660.

Wang, X., Huang, R., Sommer, M., et al. (2024). The efficacy of artificial intelligence-enabled adaptive learning systems from 2010 to 2022 on learner outcomes: A meta-analysis. *Journal of Educational Computing Research, 62*(6), 1348–1383.

Wang, Y., Pan, Z., & Solhi, M. (2025). L2 students' speaking skills in robot-assisted language learning: A meta-analysis. *European Journal of Education, 61*(1).

Weng, Z., & Fu, Y. (2025). Chinese TESOL teacher educators' perceptions of factors influencing teacher agency. *International Journal of Applied Linguistics, 36*(2), 1121–1139.

Wilschut, T., Sense, F., & van Rijn, H. (2022). *Personalized vocabulary learning using automatic speech recognition* [Preprint]. PsyArXiv.

Xu, H., Qi, C., Liu, T., et al. (2026). *Multi-agent learning path planning via LLMs* [Preprint]. arXiv.

Yan, L., Greiff, S., Lodge, J. M., et al. (2025). Distinguishing performance gains from learning when using generative AI. *Nature Reviews Psychology, 4*(7), 435–436.

Yang, C., Li, R., & Yang, L. (2025). Revisiting trends in GenAI-assisted second language writing: Retrospect and prospect. *Journal of Educational Computing Research, 63*(7–8), 1819–1863.

Yao, H., Gu, M. M., & Zhang, Y. (2026). Purpose profiles in informal digital learning of English: A mixed-methods study. *Journal of Computer Assisted Learning, 42*(3).

Yi, M., & Chen, M. (2025). The human touch in AI: Optimizing language learning through self-determination theory and teacher scaffolding. *Frontiers in Psychology, 16*, Article 1568239.

Zadorozhnyy, A., Lai, W. Y. W., & Lee, J. S. (2025). EFL teachers' ecological barriers to integrating informal digital learning of English. *TESOL Quarterly, 59*(4), 2009–2034.

Zaimoğlu, S., & Dağtaş, A. (2025). Teacher cognition and practices in using generative AI tools to support student engagement in EFL higher-education contexts. *Behavioral Sciences, 15*(9), Article 1202.

Zawacki-Richter, O., Marín, V. I., Bond, M., et al. (2019). Systematic review of research on artificial intelligence applications in higher education –

Where are the educators? *International Journal of Educational Technology in Higher Education, 16*(1).

Zeevy-Solovey, O. (2024). Comparing peer, ChatGPT, and teacher corrective feedback in EFL writing: Students' perceptions and preferences. *Technology in Language Teaching & Learning, 6*(3), Article 1482.

Zeng, B., Liu, M., Zhen, Z., et al. (2025). *GraphMASAL: A graph-based multi-agent system for adaptive learning* [Preprint]. arXiv.

Zhang, S., & Liu, X. (2025). Learner emotions in AI-assisted English as a second/foreign language learning: A systematic review of empirical studies. *Frontiers in Psychology, 16*, Article 1652806.

Zou, M., Liu, G., & Li, D. (2025). Beyond motivation: Modeling the predictive role of L2 pride in informal digital learning of English. *International Journal of Applied Linguistics, 35*(4), 1938–1952.

AI on the go: Technology is Reconstructing Second Language Assessment

Rabia Damla Karaçeper[1], *Gülay Kiray*[2] & *Tuncer Can*[3]

Introduction

The most common expectation about technology use in educational contexts is that it should achieve demanding teacher tasks and eventually replace certain responsibilities.

Technological developments, such as automated scoring systems, machine learning and artificial intelligence (AI), have so far reduced the reliance on teacher workforce both in language instruction and assessment, especially in L2 assessment, being programmed to define and quantify the certain features in order to predict human-like scores. The rapid measurement of computers in combination with humans' capacity to understand nuanced things produce efficient scores with consistency. The goal of this chapter is, therefore, to explain what automated scoring systems are, how they are operated and currently used, pointing out their contributions and controversies to L2 assessment. It continues with further explanations about AI-powered tools for classroom-based assessment and instruction.

[1] Lecturer (PhD Candidate, Istanbul University Cerrahpaşa), National Defence University, Faculty of Education, Department of English Language Teaching, ORCID: 0000-0001-6155-9973, *rabiadamla.k@msu.edu.tr*, Türkiye

[2] Assist Prof. Dr., Istanbul University Cerrahpaşa, Faculty of Education, Department of English Language Teaching, ORCID: 0000-0003-2045-8636, *gkiray@iuc.edu.tr*, Türkiye

[3] Assoc. Dr., Istanbul University Cerrahpaşa, Hasan Âli Yücel Faculty of Education, Department of English Language Teaching, ORCID: 0000-0001-8145-0772, *tcan@iuc.edu.tr*, Türkiye

1.The emergent need for technology in assessment history

The initial steps of technology into language assessment can be regarded as the time when an American multinational technology company, called IBM, first used computer to score multiple-choice tests in 1935 in the U.S. During the mid-1960s, driven by the need to streamline a more effective testing process, language testing agencies began to use technology in a variety of assessment stages such as evaluation of the test data, storing test items in databases and generating test outcomes reports for test takers.

Computer-based testing (CBT), which challenges the traditionally delivered pen-paper tests, enables an alternative assessment online or with computers. The Language Testing Research Colloquium (LTRC) in 1985 and the conference proceedings by Stansfield (1986) marked the beginning of notable attempts to fully understand second language computer-based testing in the 1980s. When properly designed, computer-based assessments of language skills may surpass the traditional types of assessment in terms of accuracy as well as effectiveness (Chalhoub-Deville & Deville, 1999; Chapelle, 2001). L2 computer-adaptive testing (CAT) first emerged in the 1990s with the breakthrough advancement of latent trait model application, which enables the selection and presentation of test items in a particular sequence depending on the examinee's responses to each item. In addition to the impact of such developing technologies, the growing popularity of smaller and more affordable computers have also paved the way for computer-adaptive testing (Reckase, 1974).

1.1. Item Response Theory and Rasch Modeling

Advances in Item Response Theory (IRT) and Rasch modeling have shown parallelism with the growing emphasis of language and skill-based testing which especially includes polytomously scored items. IRT is a collection of mathematical models which are used to model test taker's responses to test items (Lord, 1981). The main objective of IRT is to estimate test taker's ability levels and the difficulty level of the items. Rasch Modeling, which is often thought of as a subset of IRT, is a psychometric model and prescriptive measurement methodology (Rasch, 1960). They both have two models, namely dichotomous and polytomous models. In dichotomous models, the items are only limited to two possible answers such as true /

false questions. However, in polytomous models, the items may have multiple answers as in the Likert-type scales and different scores can be obtained from "strongly agree" to "strongly disagree".

Using Rasch modeling and IRT in language and skill testing provides many benefits because they ensure a more valid and reliable assessment of an examinee's abilities by guaranteeing consistency between different administrations. Capturing the subtle details of examinees' answers, minimizing measurement errors, and, more importantly, providing fair distribution of items with different difficulty levels are regarded as the absolute advantages of such models. They also open the door for development of adaptive testing, in which test items are sequenced specifically based on the level of examinee's ability. Whenever the difficulty of test items is increased not randomly but sequentially, examinees answer more correctly. As a matter of fact, many higher education institutions and language testing organizations cooperatively worked on developing CBT and CAT systems as early as the 1990s. Examples of such universities and companies include the Montgomery County Public Schools Project, Brigham Young University CAT Instruments for placement tests of French, German, and Spanish, CommuniCAT and CBT BULATS by the University of Cambridge Local Examinations Syndicate (UCLES), and the CBT Test of English as a Foreign Language (TOEFL) by the Educational Testing Service (ETS) (Phuong, & Volker, 2022).

1.2. Online Language Proficiency Tests

The widespread accessibility of computers and the Internet has led to major changes in the way to assess proficiency in language (Alderson, 2000) as to delivering and scoring the tests. For example, the growth of the Internet has propelled the developments for Web-Based Testing (WBT), where an examinee could access the test remotely on his or her computer through a website located on the examinee's server. The increased convenience of the ways to take the language tests inevitably has transformed the language testing procedures and thus, facilitated greater participation rate.

Furthermore, the concept of automated scoring has gained much attention because they could score hundreds of features and offer feedback-rich test results simultaneously in a timely manner. Specifically designed and

developed for writing and speaking skills, automated scoring systems have been utilized in order to minimize the bias and prejudice towards the test takers, who may have a wide variety of backgrounds due to the globally occurring immigration, mobility and workforce. In order to standardize initiatives in education for each and every language learner, as in line with the Common Core State standards in the United States, automated scoring systems can offer impartial assessment, rapid language-based testing, and effective scoring and reporting.

Currently, many language tests are available as online assessments worldwide. Sponsored by the Council of Europe, the DIALANG project provides diagnostic assessment in 14 languages. More examples of such internationally recognized and accepted language proficiency tests by educational institutions include TOEFL IBT by Educational Testing Service (ETS), IELTS Indicator by the British Council, Duolingo English Test, Pearson Test of English (PTE), and Linguaskill by Cambridge. These tests offer online language proficiency assessment showing minor differences in their duration, target audience or purposes. For example, TOEFL or IELTS is mostly a requirement for students when applying for a university in an English-speaking country. On the other hand, if English is needed for a more professional purpose in specific contexts like workplaces or business environments, the alternative tests are The Test of English for International Communication (TOEIC) by ETS, LanguageCert International ESOL, International Test of English Proficiency (iTEP) and Oxford Test of English.

It is evident that foreign major testing agencies and institutions have a long history of being involved in and developing various online or computer-based language proficiency assessments. In Türkiye, the ongoing situation has not been very heartwarming as language learners have had to take paper-based, multiple-choice item-type tests (YDS or YÖKDİL by ÖSYM) for many years for both academic and professional purposes. The test includes 80 questions full of mechanical multiple-choice items of grammar, vocabulary, sentence completion, translation, reading comprehension, paragraph completion, restatement, dialogue completion and irrelevant sentence omission from a paragraph. It lacks authentic language skill assessment because it does not include the testing of listening, writing and speaking skills. The same test has been delivered online, called e-YDS, since

2014, however it still fails to assess skill-based language testing. Quite recently developed, E-TEP (Electronic Test of English Proficiency) is a noteworthy advancement of online language testing from the Turkish context, which can be regarded as a revolutionary instance not only because it's delivered electronically, but also because it is designed to assess examinee's four skills of English proficiency in one compact test.

2. Automated Scoring of Writing

There are two terms commonly used to refer to computer scoring of test taker's writing and feedback to their answers in high-stakes and low-stakes tests: automated essay scoring (AES) and automated essay grading (AEG). Automated writing evaluation (AWE), on the other hand, is an umbrella term that encompasses both AES and AEG, and it outlines a more formative, feedback-rich integration of the technology into language learning. Developed in the 1960s, automated writing evaluation (AWE) systems have been employed by many language-testing organizations to enhance the effectiveness of language instruction and assessment. These systems rely heavily on a combination of statistical techniques and linguistic strategies from a variety of fields such as linguistics, computer science, and educational measurement. They investigate related linguistic evidence in test taker's writing and then create an algorithm which enables mapping the prepacked collection of linguistic input to produce a comparable score to that of human raters.

Automated essay scoring models have two types: *prompt-specific* and *generic*. In prompt-specific models, test takers from the target population are first administered to each prompt or essay topic in order to gather a sample of responses that should encompass every possible aspect of language performance or ability. The main reason for collecting such samples in advance is to create a huge corpus of relevant prompts or essays that can be utilized for modeling when scoring the incoming ones. For example, being a widely-used scoring system, Intelligent Essay Assessor®, typically requires 300–400 responses per model (Van Moere & Downey, 2016). In this way, more accurate evaluation of deeper concepts, such as essay content, task completion, organization and coherence of opinions, appropriate register use and authorial voice, becomes within the realms of *possibilities*.

Compared to prompt-specific models, generic models are only trained once and then applied to all essay prompts with the same feature set and feature-weighting strategy. In the automated scoring contexts, features are the certain characteristics of the essay that help the model to evaluate and score the test taker's writings on a pre-set scale. These features can be listed as grammar, vocabulary, sentence structure, coherence and content relevance, typically focusing more on the exterior elements, such as frequency of grammar and punctuation mistakes, amount and position of discourse markers, metrics of sentence complexity, word frequency, and so on. In the training process of the model, each feature is given weight that accounts for a proportional score of the overall essay score. For instance, if use of rich vocabulary is prioritized among the other features for the quality of the essay, it is determined with more weight than the other less important features such as the length of sentences.

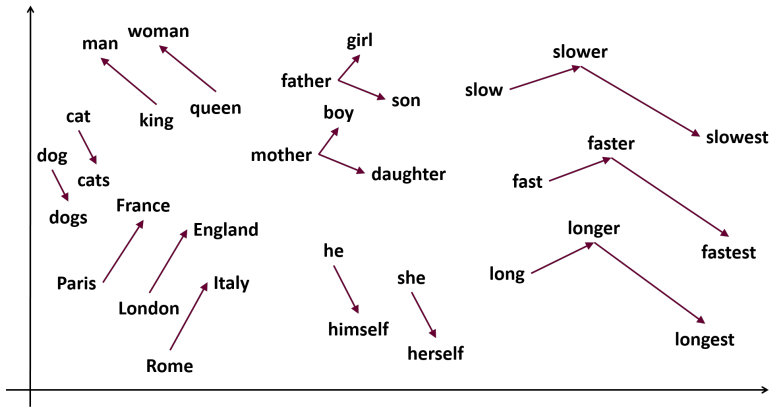
2.1. The development of AWE systems

The earliest known AES system is the Project Essay Grade developed by Ellis Page in the 1960s. At that time, the handwritten essays used to be typed into a mainframe computer manually to quantify the predefined surface linguistic elements (e.g. the length of the essay, average word length, part of speech, counts of punctuation and function words such as prepositions and pronouns, lexical diversity, etc.) of the essays without any reference to its semantic content. PEG was able to generate a holistic score correlation with human scores, with a maximum $r= 0.87$, by using multiple regression (Page, 2003).

Although the perks of PEG were groundbreaking at that time, it was unable to precisely capture the intricate nature of human writing due to the lack of in-depth linguistic and contextual assessment. This started the journey of more modern techniques for natural language processing (NLP) among which Latent Semantic Analysis (LSA) by Foltz et al. (2013) stands as a great breakthrough. LSA makes use of mathematical algorithms to identify the underlying semantic structures within enormous text corpora and to allow for a deeper comprehension of the content and meaning of the human writing.

LSA treats the written text as vectors which are interrelated within a multidimensional semantic space (Van Moere & Downey, 2016). The texts are generally vectorized at word, sentence and passage levels so that machines can process scoring through vectors, mathematical representations of the texts. At word level, every word has a vector representation and if two words occur in similar contexts, it means that they have similar vectors. These vectors depict a clear image of how words travel to some distance within the boundaries of their contextual meanings (see Figure 1). Some popular word vector techniques include *Word2Vec*, *GloVe*, and *fastText*.

Figure 1: Illustration of Word2vec Model



Similar to words, vectors can also represent sentences by combining word vectors within the sentence in relation to high-dimensional vector space. *Sentence- BERT* (SBERT) can be shown as an advanced technique to carry out such a complex relational analysis. If sentences follow each other in a logical order, the sentence vectors position in close proximity (cosine). The same holds true for the paragraph level, at which lengthy passages or documents can be vectorized, as well by using sentence vectors or applying models, such as *Doc2Vec,* to operate with larger proportions of texts. In conclusion, the effective use of word, sentence and paragraph vectors in an essay or writing has enabled machines to approximate the complicated scoring of writing skill to that of human raters. These vectors help predict

the of the coherence of writing and thus increase the reliability and validity of the automated scoring systems. Now, let's look at the two sentences below and analyze them through LSA.

> Homework is assigned weekly by a group of teachers.
>
> Every week, a team of instructors distributes assignments.

Despite the fact that these sentences share no words in common, their meanings are approximately the same depending on the contexts of the words that constitute them. For example, the words "teachers" and "instructors" appear in similar contexts and they match with each other because they mean the same thing. Since the underlying meanings of these two sentences are the same, they would be scored the same in LSA vector space. Here is the list of 14 more sentences, produced by ChatGPT pre-trained to generate more sentence vectors using LSA.

> "The school's educators allocate homework on a weekly basis.
>
> Weekly, a panel of teachers issues homework to students.
>
> Assignments are handed out each week by teaching staff.
>
> Instructors assign homework to students every week.
>
> The educators provide weekly homework tasks.
>
> Homework is given out by a collective of school teachers every week.
>
> Each week, teachers assign homework to their students.
>
> A group of school instructors assigns homework on a weekly schedule.
>
> Teaching team assigns homework to students on a weekly basis.
>
> Each week, a cadre of educators hands out tasks.
>
> Weekly, a collection of mentors provides student exercises.
>
> A set of school faculty members issues weekly assignments.
>
> A committee of lecturers allocates homework on a weekly timetable.
>
> On a weekly basis, assignments are distributed by a consortium of instructors."

Figure 2: Illustration of Sentence-BERT architecture

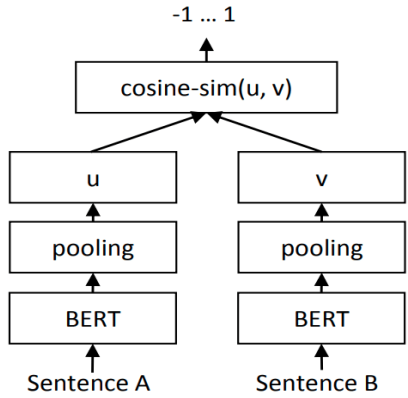
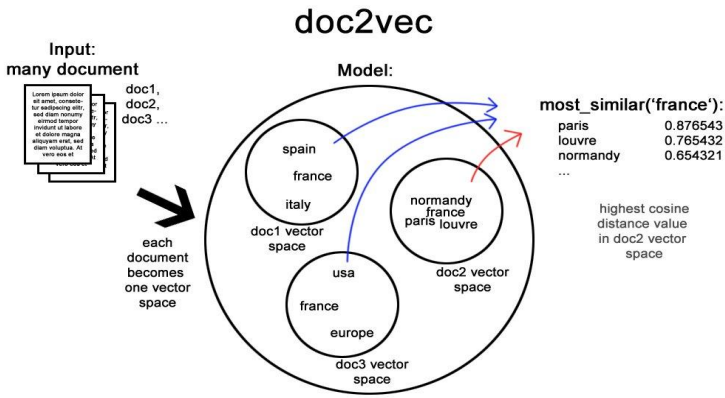


Figure 2 explains the process of two sentences being calculated for similarity through BERT models. Two sentences are put in the models to analyze the contexts of the words in a sentence and create vectors to be transformed into the pooling where fixed-sized vectors are determined, u and v. The cosine similarity between u and v is calculated with a value of -1, 0 and 1. A value of 1 indicates that the vectors are the same whereas -1 signifies that the vectors have opposite meanings. The vectors bear no similarity if the value is 0.

Figure 3: Doc2vec Model



The transformation of documents into vectors is grouped into different semantic vector spaces pertaining to a multidimensional space using cosine similarity as shown in Figure 3. All of the documents which are provided into that model contain text data, both sentence and word level, that can be clustered and positioned in accordance with their semantic similarities and differences. For instance, the word "France" might be found to be similar with vectors such as "Paris", "Louvre" and "Normandy".

2.2. Currently-used AWE systems

Currently, major testing agencies and institutions employ automated writing assessment (AWA) systems that provide consistent, immediate feedback and impartial scoring. Prominent examples include *E-rater*® and *Criterion* by ETS, *Intelligent Essay Assessor® (IEA)* and *WriteToLearn* by Pearson, and *IntelliMetric*™ and *MyAccess!* by Vantage Learning.

E-rater®, developed by Educational Testing Service (ETS), is a highly reliable automated scoring system used in high-stakes assessments such as TOEFL. It evaluates surface features of writing, including grammar, usage, mechanics, style, and organization. The system is regularly updated in line with current educational standards and linguistic research to maintain accuracy across different testing contexts.

Criterion, another ETS system, is primarily designed for formative assessment and writing practice in educational settings. It provides immediate, detailed feedback and automated scores by using e-rater® technology to evaluate the surface features of essays.

Intelligent Essay Assessor® (IEA), developed by Pearson, focuses more on the semantic content of essays through latent semantic analysis (LSA). Unlike systems that mainly assess surface features, IEA evaluates content and coherence by comparing submitted essays with previously scored essays in semantic space. It calculates the vector cosine between the new essay and pre-scored essay vectors to generate holistic scores. Approximately 300–400 pre-scored essays are required to score an incoming essay, making IEA useful for both assessment and feedback generation.

WriteToLearn, also developed by Pearson, is an adaptable educational tool widely used in K–12 settings. It uses IEA for scoring and feedback and

supports additional skills such as reading comprehension, vocabulary development, and summarization.

IntelliMetric™, developed by Vantage Learning, applies AI and machine learning to evaluate multiple writing traits, including content, organization, and mechanics. Continuously trained on new data, it has been used in more than 3 million assessments for formative and summative purposes. Its adaptive nature enables it to respond to evolving language and writing practices while maintaining robust and precise scoring.

MyAccess!, another Vantage Learning system, relies heavily on IntelliMetric™ to provide immediate scores and individualized feedback. It is widely used in schools and educational settings to support writing instruction and facilitate the teaching and learning process.

3. Automated Scoring of Speaking

The term Automated Speech Recognition (ASR) refers to the transformation of spoken responses into written text by sophisticated systems and algorithms. Early examples of such systems include *EduSpeak*, *Dragon NaturallySpeaking* and *Fluency Tutor for Google (Texthelp)*. Those systems could basically analyze speaker's features of pronunciation and correct pronunciation errors such as *Carnegie Speech's NativeAccent*; provide corrective feedback on pronunciation e.g. *Tell Me More*; and provide feedback on pronunciation in a variety of languages like *Rosetta Stone.* However, modern systems have been called for in order to evaluate a broader range of competencies along with providing nuanced analysis for widespread use.

Sophisticated models of automated scoring of speaking incorporate machine learning and advanced learning models enhance reliability and accuracy of the speaking assessments since "these systems must evaluate *what* is said as well as *how* it is said" (Van Moere & Downey, 2016, p.347). These systems can provide precise assessments of coherence, content relevance and overall communicative competence. They can also deliver elaborated analysis of various components of speaking proficiency such as fluency, lexical and grammatical complexity and discourse structure as well as pronunciation. These systems are currently being employed by well-

known testing agencies (e.g. ETS and Pearson) in many high-stakes tests, including TOEFL and PTE Academic.

With the revolutionary integration of Artificial Intelligence (AI) and machine learning models into automated scoring of speaking, these systems have developed from simple speech recognition to the evaluation of more extensive features of speech. Three crucial models must be developed in order to assess speech for these systems: acoustic model; language model; and scoring model. The utilization of these three models and properly designed tests and tasks produces accurate, consistent and rapid scoring for speaking proficiency.

3.1. Acoustic Model

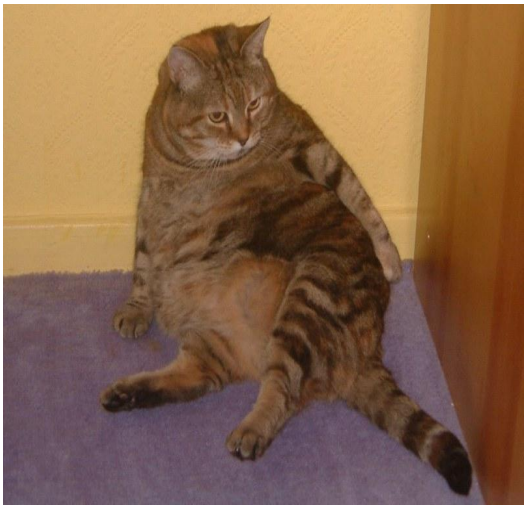
The main constituent of speech recognition is the acoustic model. The acoustic model functions converting the audio or speech first into smaller units of phonemes and then sequencing them evenly to match the speech to be recognized. These models attach great importance to the identification of key features of speech such as pitch, duration, intensity, intonation, fluency (pauses, speech rate) and pronunciation (phonetic accuracy, stress patterns) (Franco et al., 2010). Speech recognition software generally administers Hidden Markov Models (HMMs) to represent each physical sound (Young 2001). Recognition can be regarded as a set of probabilities among which the model detects the likeliest phone or word from each 10-millisecond speech frame extraction of spectral traits. Acoustic models need "training" or optimizing on a database of spoken speech by "pairing the audio speech with transcriptions of that speech, so that the model associates sounds with orthographic representations" (Van Moere & Downey, 2016, p.347). The model can be trained by any speech data such as radio broadcasts or audios from Youtube (Hinton et al. 2012).

3.2. Language Model

Following the transcription of spoken data, the language model is employed to examine the usage of grammar and vocabulary as well as the coherence of the spoken response. Through the natural language processing (NLP) techniques, the language model evaluates the appropriateness and complexity of language with regard to its syntax and semantics. The model takes on frequencies for n-grams in order to anticipate what the speaker might say.

For example, if a speaking task is to describe a picture of a cat sitting on a mat, the bigram "the cat" and the trigram "the cat sat" might appear more often in the examinee's responses (see Figure 5). Therefore, it is of vital importance to train the language model with the likeliest sequence of words pertaining to the specific task. Even though performing tasks such as reading aloud or repeating the sentences are not communicative, they are valid from a psycholinguistic perspective and they contribute greatly to the training and evaluation of automated speech recognition algorithms (Van Moere, 2012). These tasks yield more than 90% accuracy in word recognition, even with accented speech (Balogh et al. 2012). On the other hand, it gets much more difficult to recognize the words occurring in a spontaneous speech. According to Ivanov et al. (2016), the rate of word error in spontaneous non-native speaker speech is 23%, which can be reported as roughly three out of four words recognized correctly. As for the spontaneous speech from native speakers, Hinton et al. (2012) states that the optimal threshold for word recognition under ideal circumstances and adequate computing power can exceed 80%, based on the speech to be recognized.

Figure 5: *Illustration of the description of a speaking task using N-grams*



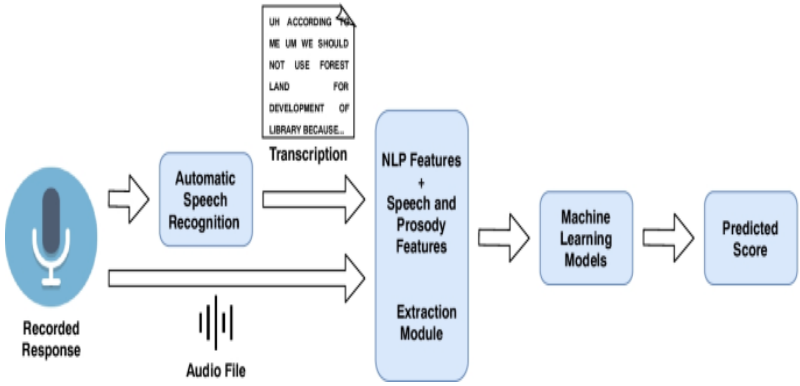
Unigrams (N=1): • 'The' • 'cat' • 'sat' • 'on' • 'the' • 'mat'	**Bigrams** (N=2): • 'The cat' • 'cat sat' • 'sat on' • 'on the' • 'the mat'
Trigrams (N=3): • 'The cat sat' • 'cat sat on' • 'sat on the' • 'on the mat'	**4-grams** (N=4): • 'The cat sat on' • 'cat sat on the' • 'sat on the mat'

3.3. Scoring Model

The ultimate step of automated scoring of speaking is the scoring model, which is known as the process of selecting extracted features both from the acoustic model and language model to predict human ratings (Van Moere & Downey, 2016). For efficient scoring, these selected features (such as

fluency, pronunciation, lexical diversity, grammar and discourse coherence) are combined to create a holistic depiction of the speech. Following that, deep learning and other advanced machine learning are run to predict scores, utilizing the training data which have been formed by the pre-scored speech by human raters. These algorithms are taught how to mimic the patterns of human scoring. Upon approximation to human scoring, it is important to include both the features of the output which are the best indicators of human scoring and predictors which are relevant to the human scoring. For example, if the presence of pauses in an examinee's spoken response weren't considered as an essential predictor of human scoring of grammatical competence by human raters or the rating scales, it wouldn't be reasonable to include this parameter in the scoring model. Figure 6 demonstrates the comprehensive timeline of how automated speech scoring systems function.

Figure 6: *Timeline for automated speech scoring system*



3.2. Commonly-used ASR systems

Automated speech recognition (ASR) systems have recently been developed for spoken language assessment. Leading examples include *SpeechRater*™ by ETS, used for English language assessment, and *Versant*™ by Pearson, which assesses languages such as Arabic, Dutch, French, Spanish, and English. *The Duolingo English Test (DET)* is a more recent example, offering affordable,

innovative, at-home language proficiency testing through artificial intelligence and automated scoring.

SpeechRater®, developed by the Educational Testing Service (ETS), is a trusted automated speech-scoring system, particularly used in the TOEFL IBT to assess English proficiency for academic purposes. Using natural language processing and acoustic modeling, it evaluates fluency, pronunciation, coherence, grammar, and vocabulary. Its scoring is based on a large, trained dataset of pre-scored responses and is aligned with human ratings.

Versant™, developed by Pearson, uses advanced speech recognition technology to provide rapid and accurate scoring in several languages, including English, French, and Spanish, across government, business, and education. It evaluates pronunciation, fluency, syntax, and lexical diversity, while minimizing the influence of accented speech.

As a more recent assessment tool, *Duolingo English Test (DET)* is known for its affordability, practicality, and innovative use of artificial intelligence in assessing speaking and writing. It has been increasingly accepted by higher education institutions for admission and can also be used for self-assessment, offering a user-friendly interface and the convenience of taking the test at home.

4. Contributions and Controversies of Technology

This section discusses the contributions of technology to L2 assessment as well as some critical issues concerning its use in L2 assessment.

4.1. Practicality

Technological advancements such as integration of artificial intelligence (AI) and machine learning in L2 assessment have contributed to an effective and affordable administration, reporting and scoring. As a result, language testing agencies and examinees, on a vast scale, benefit from this convenience temporally and financially. During the outbreak of the COVID-19, the rapid integration of video-conferencing tools such as Microsoft Teams and Zoom and virtual environments have become essential to maintain the assessment process properly. These tools allow

stakeholders to perform simultaneous and ubiquitous assessment not only by saving time and money but also by increasing easy attendance without travel. It is demanding to develop such technologies as automated scoring systems in terms of time dedication and financial investment; however, it gets quite easy to score written and spoken responses effectively in the long term. Once employed, automated scoring systems of writing and speaking tend to become budget-friendly and practical because it reduces the expenses of human raters such as hiring and training them to specialize in rating.

The practicality of delivering such tests online and on digital devices, however, may pose threats to the test security. The same technology now may be used to cheat on the tests or hack these systems. The common use of technological devices such as smart watches, electronic recorders and screen-capturing apps might be misused for the intention of cheating on tests.

4.2. Reliability

Automated scoring systems offer standardization and consistency in test administration and scoring through automated scoring of examinee's written and spoken responses. These systems increase the reliability of tests by rating consistency of scoring, unlike human raters who may be a source of inconsistency in scoring (Barkaoui, 2010). Research on AWE systems shows that correlations between the scores of human raters and an AWE system, for example Intelligent Essay Assessor (IEA), are very similar or even higher than the correlations between scores of two different human raters (Foltz et al., 2013). Moreover, automated scoring systems provide a wide variety of instructional feedback in addition to a summative assessment score on the test. The same conclusion is true for another AWE system, namely e-rater®, which is used as a complementizer to human scoring in several writing tests, such as TOEFL IBT (Attali & Burnstein, 2005). Thereupon, automated scoring systems render the maintenance of test reliability possible across items and over time. However, the reliability of the test cannot be only achieved by proper scoring. It is also important to take some crucial factors that highly affect test administration into consideration such as environmental circumstances (comfort, noise etc.), necessary equipment (e.g. calibrated devices and materials) and human administrators

(proctors' behaviors and competence). Thanks to technology, the effect of such factors can be minimized, promoting test fairness among test takers and increasing the validity of the test and its results.

4.3. Authenticity

Besides practicality and reliability, technology also offers benefits the development and use of authentic and creative tasks in L2 assessment. Test authenticity increases with the combination of different modes and becomes similar to real-life situations where speakers depend on both verbal and non-verbal clues. For example, research confirms that integrative listening tests with visuals and videos correspond more to real-world listening tasks when compared to the listening tests that are made up of solely aural input (Ockey, 2007; Wagner, 2010). Likewise, integrative speaking tests can be prepared through audios or videos that can be used as stimuli similar to real-life interactions where understanding the visual clues and non-verbal signals is generally essential in the context of authentic L2 speaking. Thus, integrative speaking and listening tests with the use of authentic modes help deliver a robust and realistic medium for assessing productive skills. Douglas (2013) lays great emphasis on the significant ability of such new technologies in that they can be designed optimally to assess test takers' language proficiency in authentic tasks and situations. As an example, interactive virtual tasks are of great importance primarily in the assessment of English for Specific Purposes in sectors such as health, business, tourism etc. It could turn out to be beneficial to prepare replicated real-world challenges for aviation English test takers and assess their performance on such virtual environments (Park, 2018). Through virtual environments, one can comfortably practice a variety of challenging tasks, for example how to manage job interviews, solve job-related problems and navigate angry conversations, with valuable instruction and detailed feedback designed to offer real-life scenarios.

4.4. Construct

Another major benefit of technological developments is closely connected to the construct of the tests, and therefore, the authenticity of the tests, as well. With the authentically designed speaking proficiency tests, it is now easily possible for examiners to measure many undiscovered areas of the language

ability, such as interactional competence, pragmatic competence or multimodal communicative competence in a computer-based testing. The use of video-conferencing tools, virtual environments, chatbots offer more comprehensive and concurrent assessment of such competences rather than traditionally delivered face-to-face speaking tests. Automated scoring systems (e.g. IntelliMetric™, SpeechRater, e-rater®) take this point to even beyond human subjectivity, by analyzing hundreds of aspects of language simultaneously (Schultz, 2013; Chen et al., 2018), which is definitely a demanding job for human raters because of multitasking and technology fatigue (Knoch et al., 2007). Given all these scenarios, technology definitely promotes the construct of the language ability to be completely implemented in tests.

Despite countless benefits for the construct of language ability, some scholars have called for the need to redefine the term construct from a broader perspective in technology-mediated tests. For example, the use of multiple modes has the potential to impede the assessment of listening ability as it may include visual elements to fully understand the aural input (Wagner, 2008). Thus, it causes a more complicated construct to measure, posing a direct threat to the validity of the test (Douglas & Hegelheimer, 2007). It is still questioned whether speaking and listening tests should be integrated or discrete, limited only to the construct of the ability that is being measured. Given this critical issue, it becomes more important to detect potential construct-irrelevant variables resulting from the use of technology, such as typing proficiency of the examinees in a writing test, to be excluded in technology-enhanced language tests.

4.5. Washback

The integration of technology in L2 assessment has inevitable impacts (called test washback) both on learning and teaching because of their exceptional abilities. To start with the learner perspective, it encourages learning among learners making them responsible for their own learning process, offering customized feedback on and scoring of learners' language proficiency. This helps learners to evaluate their own learning by making decisions about the given feedback, which becomes assessment *as* learning (Dann, 2014). Almost all of the automated scoring systems, as mentioned in the previous sections, provide detailed formative feedback and instant scoring on students' writing and speaking abilities, which boosts learner autonomy so that learners can improve their writing skill by revising their

drafts accordingly or practice independently by receiving various types of linguistic feedback provided on their speaking proficiency.

From the perspective of language teachers and testers, the integration of technology, as the phrase goes, is a savior in terms of time and labor. Automated scoring systems, as long as used in combination with human-touch such as the pedagogical expertise of subject matter experts (SME), promise to reinforce learning-oriented assessment, which refers to the concepts of assessment *of* and *for* learning (Turner & Purpura, 2016). Finally, new technologies allow for student engagement in the assessment process, making room for teachers to grant more time on the planning and content of teaching, without wasting time and energy on assessment.

More importantly, the new paths of testing language proficiency yield to unpredictable changes, in the long run, regarding both L2 teaching and learning practices. It should be the focus of future research to investigate how language teaching and learning methods will change if tests are accessible through mobile phones, computers or tablets and speaking tests are delivered through teleconferencing apps, just like making a call. How will the test preparation be for technology-based standardized tests? And more importantly, what about the behaviors of the teachers and test administrators? Therefore, the impacts of technology-integrated tests should be fully examined for all stakeholders in second language assessment, addressing adaptive educational practices to find responses to these questions. This will unquestionably secure the place of technological innovations in L2 education, which creates "opportunities to learn from both the process and the results of test taking" (Chapelle & Voss, 2014, p. 121).

5. Innovative Platforms and AI Tools for Classroom Use

Technology offers countless conveniences for language classroom use. It has been already discussed that the AWE and ASR technologies support learner autonomy with a wide range of diagnostic and formative feedback on students' writing and speaking performances with reliable and rapid scoring. This would ease the teachers' job inside and outside of the classroom, paving the way for more effective planning of the teaching content. Recent technological innovations also offer cutting-edge facilities for

classroom use, so the only thing left for teachers is to keep up with these innovations and integrate them into the classroom for efficient and practical classroom assessment. For example, web-based platforms such as *Extempore, Learnosity, Questionmark* and *Socrative* are used for educational purposes such as item writing, item bank management, test design as well as interactive task design and language practice for all areas of language learning. They provide individualized feedback and detailed reporting through versatile and interactive question types. Although it lacks depth in assessment of all components of language, *Kahoot* is a highly popular tool among students due to its engaging interface and availability.

Other technological efficiencies, such as *VoiceThread* and *Flip,* can be used for speaking formative assessment offering a more practical classroom assessment through multiple modes such as images, voices, videos and presentations and, more importantly, authentic and individualized tasks. For example, *Flip, Miro* and *Padle*t can be good e-portfolio sites for reservation of such technology-mediated tasks, so that they could be later used for formative assessment (Villares & Carciu, 2023).

When used for classroom-based speaking practice, chatbots and other dialogue-based systems in games and virtual environments may be appealing and beneficial for students' speaking skills. Forsyth et al. (2019) discovered that students mostly gave positive feedback about their multimodal dialogue system designed to assess speaking in academic contexts. For example, chatbots such as *SmallTalk2Me*, *TalkPal AI* and *ELSA AI* enable students to practice English in various real-life conversations, either for high-stakes exam or everyday English, with immediate feedback on students' performance. In addition to speaking, there are also other artificial intelligences that can be used to design authentic tasks for listening, reading, writing, grammar and vocabulary. *MagicSchool, Diffit* and *Twee* can be shown as good examples that help creating interactive lesson plans and worksheets as well as assessing comprehension at different levels.

Unlike traditional forms of teaching literacy, the integration of artificial intelligences through the use of multiple modes and medium supports learning of multiple literacies while language learning, which is called "multiliteracies" by the New London Group (1996). The use of AI-powered tools in the classroom engages students with designing meaning through multiple

forms such as texts, images, spaces, objects, sounds, speeches etc. These tools help students to experience the digital resources, receive explicit instruction on their effective use, offer social and critical understandings and create their own product at the end. Some useful AI-powered tools with unique capabilities include examples such as *Gamma AI* to deliver quick professional presentations, *Microsoft Copilot* to assist writing and editing, *Rewordify* to simplify the language according to the proficiency levels with focus on vocabulary building, *Thispersondoesnotexist* to create fake people faces, *Lalamu Studio (*now on Canva) to make characters talk synchronizing lip movements with created video characters, *AI Debate* to promote critical thinking, *Sumo* to produce songs from different genres, *Podcastle* to produce podcasts and *Storyboarder* to write scripts for creative story.

There is no doubt that technology, particularly AI, boosts a variety of 21st century competencies which are also known as the "4Cs" - creativity, critical thinking, communication and collaboration (Thornhill-Miller et al., 2023), making assessment an entertaining part of learning. Therefore, it is proven that appropriate and responsible use of these technologies furnishes students and instructors with comprehensive knowledge of language and required mastery of digital tools even in simple classroom-based assessment.

Table 1: List of applicable AI-powered tools for classroom-based assessment

Name	**Type of AI**	**Focus Skills**	**Capabilities**	**Advantages**	**Disadvantages**
MagicSchool	Generative AI	Reading Writing Listening	Creating lesson plans and materials Writing individualized education programs	Content can be translated into more than 100 languages Has 60+ AI tools	Limited to pre-made AI templates
StoryBoarder	Generative AI	Reading Writing	Generating accelerated video-script Creating visual representation	Improves collaboration and communication	Learning curve

Diffit	Generative AI	All areas of language	Creating distinctive educational materials Generating lesson plans and worksheets	Offers printable or digital materials	Requires editing Depends on AI accuracy
Twee	Generative AI	Reading Grammar Vocabulary	Generating micro lessons Creating quizzes and flashcards	Saves time by offering short-cuts	Designed only for English teachers
Gamma	Generative AI	Writing Speaking	Using rich content of materials Creating presentation, documents or webpages	Offers chatbot-based dialogues and multiple languages	Requires content editing and rewriting
Rewordify	Discriminative AI	Reading Vocabulary	Producing simpler texts Improving reading comprehension	Offers definitions for in-text words	Limited to text
Podcastle	Generative AI	Listening Speaking	Generating podcasts	Works with Spotify	Dependent on AI quality
Suno	Generative AI	Listening Speaking Writing	Generating songs	Offers diverse genres of music with instrumentation	May include pronunciation errors
ThisPersonDoesNotExist	Generative AI & Discriminative AI	None	Generating realistic faces and portraits	Can be used as avatars or characters	Causes ethical considerations Limited to face generation
Lalamu Studio	Generative AI	Listening Speaking	Creating lip-sync videos	Offers text-to-speech	Limited to English and German languages

				pro-cessing	
Microsoft Copilot	Genera-tive AI	Writing	Creates texts from differ-ent genres	Offers a notebook Suggests editing Checks grammar	Lack of suf-ficient con-text
ELSA AI	Genera-tive AI	Speaking	Offers real-istic role-plays with personalized feedback	Has an AI voice Detects pronunci-ation er-rors	Poorly de-signed main dashboard
Small-Talk2Me	Genera-tive AI	Speaking	Analyzing features of speech through real-istic dia-logues	Enables voice re-cording Provides a partner for mock exam	Evaluation errors of pronuncia-tion
Talkpal AI	Genera-tive AI	Speaking Listening Reading Writing	Offers Im-mersive con-versations with real-time feed-back	Offers more than 55 lan-guages	Lack of hu-manized feedback Lack of structured lessons
AI Debate	Genera-tive AI	Speaking Writing	Offers simu-lated debates Facilitating interactive discussion	Promotes critical thinking	Dependent on AI qual-ity
VATTS AI Buddy	Genera-tive AI	Reading Writing Speaking Listening	Promotes educational play Develops so-cial, cogni-tive and emotional skills	Serves as a virtual AI buddy for kids	Limited to kids aged 3 and above

Conclusion

The central idea of this chapter is that technology has revolutionized the assessment methods of L2 proficiency with the development of automated scoring systems, digital platforms and different types of generative AI powered tools, providing individualized feedback, instant reliable scoring and in-depth analysis through authentic test designs. The contributions of technology to the L2 assessment come with some possible risks, such as test security, irrelevant-construct variables and unpredictable impacts on teaching and learning regarding the delivery of the assessment. Therefore, test developers should ensure the safe delivery of the tests and make sure the construct is clear enough to measure what is aimed to be measured. On the other hand, technological advances have improved the availability, reliability, accuracy, authenticity, objectivity, and delivery speed of the tests (Van Moore & Downey, 2016), which proves that the impartial and transparent methodologies are imperative for equitable and valid assessment. However, human touch is still needed based on the evidence from empirical research and limitations of current technologies which are AI dependent. They should be used in alignment with teachers' instruction and humanized feedback for classroom-based assessments in order that students' reliance on teachers decreases and students can improve their skills autonomously. AI-powered tools allow for creation and utilization of authentic and interactive tasks both for classroom practice and classroom-based assessment. These tools still need to be perfected and their impacts should be investigated through empirical research. All in all, technology will continue to thrive, playing a determining role in the fate of L2 assessment. As it evolves, more innovative and practical assessment types will likely emerge with superior learning outcomes for language learners with diverse needs.

References

Alderson, J. C. (2000). Technology in testing: The present and the future. *System, 28*(4), 593–603.

Attali, Y., & Burstein, J. (2005). *Automated essay scoring with e- rater v. 2*. 0. (ETS Research Report No. RR- 04- 45). Educational Testing Service.

Balogh, J., Bernstein, J., Cheng, J., Van Moere, A., & Suzuki, M. (2012). Validation of automated scoring of oral reading. *Educational and Psychological Measurement, 72*(3), 435–452.

Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly, 7*(1), 54– 74.

Chalhoub-Deville, M., & Deville, C. (1999). Computer adaptive testing in second language contexts. *Annual Review of Applied Linguistics, 19*, 273–299.

Chapelle, C. (2001). Computer applications in second language acquisition: Foundations for teaching, testing, and research. Cambridge, UK: Cambridge University Press.

Chapelle, C. A., & Voss, E. (2014). 20 years of technology and language assessment in Language Learning & Technology. *Language Learning & Technology, 20*(2), 116– 128.

Chen, L., Zechner, K., Yoon, S., Evanini, K., Wang, X., Loukina, A., Tao, J., Davis, L., Lee,

C. M., Ma, M., Mundkowsky, R., Lu, C., Leong, C. W. & Gyawali, B. (2018), *Automated scoring of nonnative speech using the SpeechRater^{SM} v. 5.0 Engine*. (ETS Research Report No. RR- 18- 10). Educational Testing Service.

Dann, R. (2014) Assessment as learning: Blurring the boundaries of assessment and learning for theory, policy and practice. *Assessment in Education: Principles, Policy & Practice, 21*(2), 149– 166.

Douglas, D. (2013). ESP and assessment. In B. Paltridge & S. Starfield. (Eds.), *The handbook of English for specific purposes* (pp. 367– 383). John Wiley & Sons.

Douglas, D., & Hegelheimer, V. (2007). Assessing language using computer technology. *Annual Review of Applied Linguistics, 27,* 115– 132.

Foltz, P. W., Streeter, L. A., Lochbaum, K. E., & Landuer, T. K. (2013). Implementation and applications of the Intelligent Essay Assessor. In M. D. Shermis and J. Burnstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 68–88). Routledge.

Forsyth, C. M., Luce, C., Zapata- Rivera, D., Jackson, G. T., Evanini, K., & So, Y. (2019) Evaluating English language learners' conversations: Man vs. machine. *Computer Assisted Language Learning, 32*(4), 398–417.

Franco, H., Bratt, H., Rossier, R., Rao, R., Abrash, V., & Cesari, F. (2010). EduSpeak®: A speech recognition and pronunciation scoring toolkit for

computer-aided language learning applications. *Language Testing, 27*(3), 401-418.

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine 29*(6), 82–97.

Knoch, U., Read, J., & von Randow, J. (2007). Re- training writing raters online: How does it compare with face- to- face training? *Assessing Writing, 12*(1), 26– 43.

Lord, F. M. (1981). Applications of Item Response Theory to Practical Testing Problems. *Journal of Educational Measurement, 18*(3), 178-180.

Ivanov, A. V., Lange, P. L., Suendermann-Oeft, D., Ramanarayanan, V., Qian, Y., Yu, Z., & Tao, J. (2016). Speed vs. accuracy: Designing an optimal ASR system for spontaneous non-native speech in a real-time application. *Proceedings of the 7th International Workshop on Spoken Dialogue Systems, IWSDS 2016*, January, 12–16, 2016, Saariselkä, Finland.

New London Group (1996). A pedagogy of multiliteracies: Designing social futures. *Harvard Educational Review, 66*, 60-92.

Ockey, G. (2007). Construct implications of including still image or video in computer- based listening tests. *Language Testing, 24*(4), 517– 537.

Page, E. B. (2003). Project Essay Grade: PEG. In: Mark D. Shermis and Jill Burstein (Eds.). Automated essay scoring: *A cross-disciplinary perspective*, 39–50. Mahwah, NJ: Lawrence Erlbaum Associates.

Park, M. (2018). Innovative Assessment of aviation English in a virtual world: Windows into cognitive and metacognitive strategies. *ReCALL Journal, 30*(2), 196-213.

Phuong, N., & Volker, H. (2022). Technology and assessment. In N. Ziegler & M. González-Lloret (Eds.), *The Routledge handbook of second language acquisition and technology* (pp. 107–118). Routledge.

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests.* Nielsen & Lydiche.

Reckase, M. D. (1974). An interactive computer program for tailored testing based on the one-parameter logistic model. *Behavior Research Methods and Instrumentation, 6*(2), 208–212.

Schultz, M. T. (2013). The IntelliMetric™ automated essay scoring engine – A review and an application to Chinese essay scoring. In M. D. Shermis & J. Burnstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 89– 98). Routledge.

Stansfield, C. W. (Ed.). (1986). *Technology and language testing*. Washington, DC: Teachers of English to Speakers of Other Languages.

Thornhill-Miller, B., Camarda, A., Mercier, M., Burkhardt, J. M., Morisseau, T., Bourgeois-Bougrine, S., ... & Lubart, T. (2023). Creativity, critical thinking, communication, and collaboration: assessment, certification, and promotion of 21st century skills for the future of work and education. *Journal of Intelligence, 11*(3), 54.

Turner, C. E., & Purpura, J. E. (2016). Learning- oriented assessment in the classroom. In D. Tsagari & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 195– 208). DeGruyter Mouton.

Van Moere, A. (2012). A psycholinguistic approach to oral language assessment. *Language Testing, 29*(3), 1-20.

Van Moere, A., & Downey, R. (2016). Technology and artificial intelligence in language assessment. In Dina Tsagari & Jayanti Banerjee (Eds.), *Handbook of Second Language Assessment,* De Gruyter Mouton.

Villares, R., & Carciu, O. M. (2023). Technology-Mediated Tasks for Formative Assessment in an e-Portfolio Environment. In S. W. Chong & H. Reinders (Eds), *Innovation in Learning-Oriented Language Assessment* (pp. 219-236). Cham: Springer International Publishing.

Wagner, E. (2008). Video listening tests: What are they measuring? *Language Assessment Quarterly, 5*(3), 218– 243.

Wagner, E. (2010). The effect of the use of video texts on ESL listening test-taker performance. *Language Testing, 27*(4), 493- 513.

Young, S. (2001). Statistical modeling in continuous speech recognition (CSR). *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, 562–571. San Francisco: Morgan Kaufman.

Comparison of Traditional Writing and ChatGPT-Assisted Writing in French as a Foreign Language (FLE)

Zeynep Büyüksaraç[1]

1. Introduction

Foreign language education has become an essential component of contemporary education, as multilingual competence enables individuals to communicate across cultures and has gained increasing importance in a globalized world. Language learning therefore holds a significant place both in the French educational system and in educational contexts worldwide. Within this broader setting, French has a particularly important role. It is not only the official language of France but also one of the most widely spoken languages in the world and an official language of numerous international organizations (Ghavamian & Dogani, 2024, p. 40).

Writing is an essential language activity that supports both communication and the transmission of information. It enables individuals to express their emotions, thoughts, and experiences while also providing a means of preserving information for later use. Despite technological developments, writing continues to serve as an effective and reliable way of recording and transmitting information (Khalifa, 2020). These functions show that writing is not limited to communication; it is also a cognitive process that contributes to the organization of thought. Writing can be understood as a process in which individuals develop a text while having a certain degree of freedom in how they express and organize their ideas. From a cognitive perspective, this process involves information processing and the establishment of connections between prior and newly acquired knowledge. Writing

[1] Asst. Prof. Dr., Bolu Abant İzzet Baysal University, School of Foreign Languages, ORCID: 0000-0001-7683-0965, *zeynepbuyuksarac@ibu.edu.tr*, Türkiye

activities aim to develop language use, strengthen students' motivation to write and allow them to choose how they express themselves in written form (Moulay Omar & Hezabra, 2025, p. 68). Writing also plays an important role in students' academic and professional development. A strong command of writing helps individuals express their ideas more effectively in academic and workplace contexts and contributes to academic achievement. Writing skills are required in a wide range of practical tasks, including the preparation of formal and personal correspondence, résumés, and reports. In addition, writing can support the development of critical, logical, and analytical thinking, improve students' performance in examinations and contribute to more effective oral communication (Bora, 2023).

Technological developments have transformed the methods and tools used in foreign language teaching, creating learning environments that are increasingly interactive and learner-centered. In particular, digital tools such as Web 2.0 technologies and artificial intelligence can support the development of writing skills by making the learning environment more concrete and dynamic (Kökçü & İpek, 2025, p. 708). Artificial intelligence refers to technologies that enable computers and digital systems to perform tasks commonly associated with human intelligence, including learning, reasoning, problem-solving, and perception. In educational contexts, AI offers new opportunities to personalize learning, enhance instructional practices and improve students' learning experiences (Nguyen, 2025). With recent developments in artificial intelligence, ChatGPT became publicly available in November 2022 and quickly attracted considerable attention because of its technical capabilities and potential applications across different fields. Its basic mode of interaction is relatively straightforward: users enter a question or present a problem and the system generates a response (Courant et al., 2024, p. 75). GPT-3 is a deep-learning-based language model capable of generating human-like text, answering questions, and producing content in different languages. It was trained on large datasets collected from the Internet (Floridi & Chiriatti, 2020, p. 684). ChatGPT was also developed to respond to a wide range of user needs and to generate responses suited to different expectations, which has contributed to its use across numerous fields (Laidoudi, 2026, p. 320). The use of ChatGPT in education has also become increasingly widespread. In writing instruction, GPT-based tools may contribute to student participation and motivation (Tate et al., 2023).

In written production, ChatGPT can help students organize their ideas more effectively and support the production of clearer and more coherent texts. It can also assist with the correction of spelling and grammatical errors and thereby contribute to students' understanding of material covered in class (Tendero et al., 2024). At the same time, the growing use of large language models such as ChatGPT is changing approaches to writing instruction in both first-language education and French as a foreign or second language. Beyond generating text, ChatGPT can support different stages of the writing process, provide automated assessment and offer pedagogical feedback when requested. In the assessment of written production, for example, it can provide rapid feedback on grammar, word choice, textual coherence, and argumentative organization (Abdelghani, 2025, p. 84).

Rubrics are frequently used in the assessment of written production because they allow different aspects of students' performance to be evaluated in a systematic and detailed manner. According to Çetin (2011) rubrics are scoring tools designed to guide the assessment of different tasks. By assigning a specific number of points to each criterion, they provide evaluators with a clearer and more structured framework for examining and scoring written texts. For this reason, rubrics can serve as useful tools in the assessment of written production. One of the best-known analytic scoring tools in second-language writing assessment is the ESL Composition Profile developed by Jacobs et al. (1981). Brooks (2012) notes that the ESL Composition Profile is one of the most recognizable rubrics in second-language writing and that it was developed around five main traits for evaluating compositions produced by second-language learners. The profile evaluates writing across five dimensions: Content (30 points), Organization (20 points), Vocabulary (20 points), Language Use (25 points) and Mechanics, including spelling and punctuation (5 points).

2. Purpose and Significance of the Study

This study aims to compare the writing performance of students learning French as a foreign language under two writing conditions: traditional writing and ChatGPT-assisted writing. For this purpose, the students' compositions are evaluated using the ESL Composition Profile developed by Jacobs et al. (1981). The rubric assesses writing across five dimensions:

Content (30 points), Organization (20 points), Vocabulary (20 points), Language Use (25 points), and Mechanics, including spelling and punctuation (5 points). Each composition is scored separately according to these criteria, and the results are compared to identify differences between the two writing conditions. The significance of the study lies in its multidimensional examination of ChatGPT-assisted writing in the context of French as a foreign language. Previous research has addressed written production in FLE from different perspectives, including writing assessment, assessment tools, and the analysis of students' written texts. However, to the best of the researcher's knowledge, no study identified in the reviewed FLE literature has directly compared traditional and ChatGPT-assisted writing using the ESL Composition Profile developed by Jacobs et al. (1981). The study therefore compares 42 compositions produced by 42 students with 21 students in the traditional writing group and 21 in the ChatGPT-assisted writing group. By examining the two groups across the five dimensions of the rubric, the study provides a detailed comparison of their written performance in French. The findings are expected to contribute to research on the use of artificial intelligence in French language teaching and to provide a basis for further studies on AI-assisted writing in foreign language education.

3. Methodology

3.1 Research Design

This study employed a quantitative comparative approach to examine differences in writing performance between students who completed a writing task under traditional conditions and those who completed the same task with ChatGPT support. Comparative research focuses on identifying similarities and differences among different objects of study and may involve quantitative or qualitative empirical approaches (Lai & Fong, 2024). The written productions of the two groups were compared using the ESL Composition Profile developed by Jacobs et al. (1981). The rubric evaluates writing across five dimensions: Content (30 points), Organization (20 points), Vocabulary (20 points), Language Use (25 points) and Mechanics, including spelling and punctuation (5 points). The scores obtained from these dimensions were used to compare the writing performance of the two groups.

3.2 Participants

The study involved 42 students from the Department of Translation and Interpreting. Participants who met the predetermined inclusion criteria were selected through criterion sampling, a purposive sampling method in which cases are selected according to criteria relevant to the aims of the study (Marshall & Rossman, 2014). The main participation criteria were learning French as a foreign language, studying French using a B1-level coursebook as part of their regular instruction and participating voluntarily. The participants' level was determined on the basis of the B1-level French coursebook used in their regular language instruction. The B1 level refers to the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2022). No separate proficiency test was administered for the purposes of this study. The 42 participants were subsequently randomly assigned to two groups of 21 students, the traditional writing group and the ChatGPT-assisted writing group. Students in the traditional group completed the writing task without digital assistance, whereas students in the ChatGPT-assisted group were permitted to use ChatGPT during the writing process. All participants had prior experience with ChatGPT, while its use during the writing task was permitted only in the ChatGPT-assisted group. Among the 42 participants, 12 were male (28.6%) and 30 were female (71.4%). Most participants were between 18 and 22 years old, while a smaller number were between 23 and 27 years old. All participants reported using ChatGPT on a daily basis. Their reported purposes included translation, written production, text correction, obtaining feedback and searching for words and meanings. The most common uses were translation, writing and language-related feedback.

3.3 Research Instruments

The students' compositions were evaluated using the ESL Composition Profile developed by Jacobs et al. (1981). The rubric consists of five dimensions: Content (30 points), Organization (20 points), Vocabulary (20 points), Language Use (25 points) and Mechanics, including spelling and punctuation (5 points). Each dimension is scored separately and the scores are combined to obtain an overall writing score. Analytic rubrics allow different aspects of a product or performance to be assessed according to separate criteria. By assigning an individual score to each criterion, they

provide a more detailed evaluation of students' strengths and areas that require improvement (Çepni, 2011). In the present study, the five dimensions and scoring ranges of the ESL Composition Profile were applied to the students' written productions in French without modifying the original scoring structure. Each composition was scored by the researcher according to the criteria proposed by Jacobs et al. (1981).

3.4 Data Collection Process

Data were collected in a classroom setting. Before the writing task, students were informed about the purpose and procedures of the study and participation was voluntary. The participants were randomly assigned by the researcher to two equal groups of 21 students, the traditional writing group and the ChatGPT-assisted writing group. The data collection session lasted approximately 60 minutes. Both groups were asked to write a text on the topic "Unforgettable Holidays." The task was prepared in accordance with the B1 instructional level followed in the students' regular French course, corresponding to the B1 level of the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2022). Students were expected to produce a text of approximately 120–180 words. The same topic, time limit and length requirements were applied to both groups. Students in the traditional writing group completed the task without AI assistance, whereas students in the ChatGPT-assisted group were permitted to use ChatGPT during the writing process. Before the task, the students in this group were informed about the conditions of ChatGPT use. They were told that ChatGPT should be used only as a support tool and that their texts should not be generated entirely by artificial intelligence. They were expected to base their compositions on their own ideas, language knowledge, and writing skills. Students were allowed to use ChatGPT to generate ideas, obtain vocabulary suggestions, correct certain grammatical errors, and receive feedback on their writing. They were also instructed to complete the task individually, avoid using ready-made content from the Internet and produce an original text. All participants completed their compositions within the allotted time and the texts were collected at the end of the session for evaluation.

3.5 Data Analysis

The data were analyzed using descriptive statistics. Each composition was evaluated separately by the researcher using the ESL Composition Profile developed by Jacobs et al. (1981). Scores were assigned for each of the five dimensions of the rubric: Content, Organization, Vocabulary, Language Use and Mechanics. First, an individual score was recorded for each participant in each dimension. The participants were then classified according to the four performance levels defined in the ESL Composition Profile: "excellent to very good," "good to average," "fair to poor," and "very poor." Frequencies and percentages were calculated to show the distribution of participants across these performance levels. Mean scores were subsequently calculated for each dimension within the traditional writing group and the ChatGPT-assisted writing group. Separate tables were prepared to compare the two groups in terms of content, organization, vocabulary, language use and mechanics. Selected excerpts from the students' written productions were also used to illustrate and explain the patterns observed in the quantitative results. Finally, the mean scores of the two groups were compared across the five dimensions and an overall score was calculated to provide a general comparison of writing performance under the two writing conditions.

4. Results

This section presents the findings from the compositions produced under the traditional writing and ChatGPT-assisted writing conditions. Writing performance was evaluated using the ESL Composition Profile developed by Jacobs et al. (1981) across five dimensions: Content, Organization, Vocabulary, Language Use, and Mechanics. The results are presented in terms of mean scores and performance levels for each group. Selected excerpts from the students' written productions are also used to illustrate the patterns observed in the quantitative findings.

4.1 Analysis of the Traditional Writing Group

This section presents the results for the traditional writing group according to the five dimensions of the ESL Composition Profile. The findings are reported through tables showing score distributions and mean scores, together with selected excerpts from the students' written productions to illustrate the results.

Table 1. Content Criterion – Traditional Group

Level	Score Range	Participants	Percentage
Excellent to Very Good	30-27	P2, P4, P12, P13, P16, P20	28.6%
Good to Average	26-22	P5, P6, P7, P8, P9, P10, P15, P17, P19	42.9%
Fair to Poor	21-17	P1, P3, P11, P14, P18, P21	28.6%
Very Poor	16-13	-	0%
Total	-	21	100%
Overall Mean	-	-	23.5 / 30

The results for the content criterion indicate that the traditional writing group generally addresses the assigned topic with a basic to adequate level of development. The overall mean is 23.5/30, placing the group in the "good to average" range. Six participants (28.6%) are rated "excellent to very good," nine (42.9%) "good to average," and six (28.6%) "fair to poor." No participant is rated "very poor." Overall, most students present relevant holiday experiences and support them with some detail, although the extent of development varies across the texts. The texts in the "excellent to very good" range (P2, P4, P12, P13, P16, P20) generally contain more specific details about activities, places and experiences. P2, for example, includes *"Nous avons chanté des chansons"*, *"nous nous sommes promenés au marché"*, and *"nous avons acheté quelques souvenirs"* which add concrete actions to the account. P4 similarly includes relevant details such as *"Nous visitions des sites historiques"* and *"nous avons mangé des poissons délicieux."* P12 develops the setting through *"Des lions, des éléphants, des hyènes, des phacochères,"* while P16 names specific places and activities in *"On a visité la ville antique de Tios et le château de Filyos"* and *"on a mangé de la pastèque et des fraises ensemble."* P20 also expands the account through activities such as *"nous bronzions, nagions et jouions à différents sports"* and *"nous découvrions de nouveaux endroits amusants."* These examples show how the higher-scoring texts develop the holiday experience through specific events, places, activities, and descriptive details. The texts in the "good to average" range (P5, P6, P7, P8, P9, P10, P15, P17, P19) remain relevant to the topic but are generally less fully developed. P5,

for example, recounts seeing the sea for the first time and beginning to drown through *"Moi je jamais voir mer avant"* and *"Je commence me noyer."* These details make the event identifiable but the experience is not developed further. P9 states that *"Milena a adoré la Turquie,"* which adds a reaction to the trip rather than only listing events. Across this group, the main events are generally clear but they are supported by fewer details than in the highest-scoring texts. The texts in the "fair to poor" range (P1, P3, P11, P14, P18, P21) show more limited development. P3 mentions going to Albania through *"Nous avons allons Albenia"* and comments on the food in *"Le mangé est bien,"* but provides little additional information about the places visited, activities or events of the holiday. P1 mentions meeting friends, going to the sea, and describes the holiday as beautiful through *"Mes ami j'ai rencontré," "La mer, nous avons alle,"* and *"Ces vacances, tres belle."* These details make the general holiday experience identifiable but the account is not developed with further events or descriptions. Similar limitations appear in the other texts in this band. Overall, content performance in the traditional group is concentrated in the "good to average" range. Stronger texts develop the holiday experience through more specific events, places, activities and descriptions, whereas weaker texts provide fewer supporting details and more limited elaboration.

Table 2. Organization Criterion – Traditional Group

Level	Score Range	Participants	Percentage
Excellent to Very Good	20-18	P2, P4, P12, P13, P16, P20	28.6%
Good to Average	17-14	P5, P6, P7, P8, P9, P10, P11, P14, P15, P17, P19	52.4%
Fair to Poor	13-10	P1, P3, P18, P21	19.0%
Very Poor	9-7	-	0%
Total	-	21	100%
Overall Mean	-	-	15.4 / 20

The results for the organization criterion show that the texts in the traditional writing group generally present events in a recognizable order that allows the narrative to be followed. The overall mean is 15.4/20, placing the group in the "good to average" range. Six participants (28.6%) are rated "excellent to very good," eleven (52.4%) "good to average," and four (19.0%) "fair to poor." No participant is rated "very poor." Overall, most students organize events in a generally logical order, although the use of transitions and explicit links between events varies across the texts. The texts in the "excellent to very good" range (P2, P4, P12, P13, P16, P20) show clearer sequencing and more consistent use of expressions that connect one event to the next. P2, for example, uses temporal markers such as *"D'abord"*, *"Après"*, *"Puis"*, and *"À la fin du voyage"* to guide the reader through the sequence of events. P16 similarly uses *"Le matin"*, *"Plus tard"*, and *"Enfin"* to mark different stages of the narrative. P12 and P20 also present their activities in a clear chronological order, while P4 and P13 organize their events in a sequence that can be followed throughout the texts. The texts in the "good to average" range (P5, P6, P7, P8, P9, P10, P11, P14, P15, P17, P19) generally present events in a logical order but transitions between them are less consistent. In some texts, the sequence of events is easy to follow even though linking expressions appear only occasionally. In others, short or repetitive sentence patterns separate the events rather than connecting them into a continuous narrative. P19, for example, presents events in an understandable order but uses relatively few transitions to connect one event to the next. The texts in the "fair to poor" range (P1, P3, P18, P21) show more limited organization and fewer explicit connections between events. P3 uses sequencing expressions in *"Nous avons voyagé en voiture"*, *"Après, nous avons mangé au restaurant"* and *"Ensuite, nous sommes rentrés à l'hôtel."* These expressions establish a chronological order but the transitions between events remain limited. P1 also presents events in sequence through *"Mon ami venu"*, *"Bus nous partis"* and *"Après plage retour maison."* However, the events are placed one after another with little explicit linking between them. Overall, organization performance in the traditional group is concentrated in the "good to average" range. Stronger texts use temporal markers and transitions more consistently to connect events, whereas weaker texts rely more heavily on short statements with fewer links between them.

Table 3. Vocabulary Criterion – Traditional Group

Level	Score Range	Participants	Percentage
Excellent to Very Good	20-18	P2, P4, P12, P20	19.0%
Good to Average	17-14	P5, P6, P7, P8, P9, P10, P13, P15, P16, P17	47.6%
Fair to Poor	13-10	P1, P3, P11, P14, P19	23.8%
Very Poor	9-7	P18, P21	9.5%
Total	-	21	100%
Overall Mean	-	-	14.5 / 20

The results for the vocabulary criterion show that the traditional writing group generally uses enough vocabulary to communicate the main events and descriptions in their holiday texts, although the range and specificity of word choice vary across participants. The overall mean is 14.5/20, placing the group in the "good to average" range. Four participants (19.0%) are rated "excellent to very good," ten (47.6%) "good to average," five (23.8%) "fair to poor," and two (9.5%) "very poor." The texts in the "excellent to very good" range (P2, P4, P12, P20) show a broader range of descriptive vocabulary. P2, for example, uses expressions such as *"Séjour agréable"*, *"temps magnifique"*, and *"bons souvenirs"* which add descriptive detail to the holiday account. P12 similarly uses *"Montagnes verdoyantes"*, *"animaux sauvages"*, *"feu de camp"*, and *"aventures intéressantes"*, allowing the places, surroundings, and activities to be described more specifically. The vocabulary used in these higher-scoring texts therefore goes beyond naming basic events and includes more varied words for describing the holiday experience. The texts in the "good to average" range (P5, P6, P7, P8, P9, P10, P13, P15, P16, P17) generally communicate the intended meaning clearly, but use a narrower range of vocabulary. P16, for example, includes *"vieux bâtiments"*, *"petites rues"* and *"soirée amusante."* These expressions add descriptive information about places and experiences, but the range of lexical choices remains more limited than in the highest-scoring texts. The texts in the "fair to poor" range (P1, P3, P11, P14, P19) rely more heavily on common and repeated words. P3 uses verbs such as *"voyagé"*,

"joué" and *"regardé"*, while *"Le repas était bon"* expresses an evaluation of the meal with a simple adjective. P1 similarly uses short descriptive expressions such as *"Très joli"* and *"très heureux."* These choices communicate the intended meaning but provide fewer specific or varied words for describing people, places, activities and experiences. The texts in the "very poor" range (P18 and P21) show the narrowest lexical range in the group with limited variation in word choice and fewer lexical resources for developing descriptions. Overall, vocabulary performance in the traditional group is concentrated in the "good to average" range. Higher-scoring texts use a wider range of specific and descriptive words, whereas lower-scoring texts depend more heavily on common and repeated vocabulary.

Table 4. Language Use Criterion – Traditional Group

Level	**Score Range**	**Participants**	**Percentage**
Excellent to Very Good	25-22	P2, P4, P12, P20	19.0%
Good to Average	21-18	P6, P7, P9, P10, P13, P16, P17	33.3%
Fair to Poor	17-11	P1, P3, P5, P8, P11, P14, P15, P19, P21	42.9%
Very Poor	10-5	P18	4.8%
Total	-	21	100%
Overall Mean	-	-	17.6 / 25

The results for the language use criterion show considerable variation in grammatical performance across the traditional writing group. The overall mean is 17.6/25. Four participants (19.0%) are rated "excellent to very good," seven (33.3%) "good to average," nine (42.9%) "fair to poor," and one (4.8%) "very poor." Many students produce understandable sentences but difficulties with verb forms, auxiliaries, agreement and sentence structure remain common. The texts in the "excellent to very good" range (P2, P4, P12, P20) show more accurate use of verb tenses and more complete sentence structures. P2, for example, uses accurate passé composé forms in *"Nous avons pris le bus"*, *"nous avons regardé la mer"* and *"nous avons passé une bonne journée."* P12

also produces well-formed sentences such as *"Nous avons appris beaucoup de choses"* and *"tout le monde était content."* In these texts, tense formation and sentence structure are generally used more consistently and accurately. The texts in the "good to average" range (P6, P7, P9, P10, P13, P16, P17) are generally understandable and include appropriate tense forms, although sentence structures are often relatively simple. P9 uses *"Elle aimait beaucoup la ville,"* while P13 includes *"je regardais la mer chaque soir."* Both examples show appropriate use of the imparfait. However, the texts in this range rely more heavily on simple sentence patterns than the highest-scoring texts. The texts in the "fair to poor" range (P1, P3, P5, P8, P11, P14, P15, P19, P21) contain more frequent problems with verb formation and sentence structure. P3 includes *"Nous avons partir hier"*, *"Je voudrais retourné"*, and *"Nous mange beaucoup,"* showing difficulties with the formation of verb phrases, verb endings, and tense use. P5 similarly uses *"Je suis tombé dans mer"* and *"Moi jamais allé vacances."* In P1, expressions such as *"Moi rencontré amis"* and *"Nous aller plage"* communicate the intended meaning, but the sentences lack elements required for complete grammatical structures. P18 is the only participant in the "very poor" range with a score of 9. Forms such as *"Nous resté dans hôtel"* and *"J'ai acheté beaucoup cadeau"* show repeated problems with auxiliary use, agreement, and sentence construction. Overall, language use presents greater difficulty for many students in the traditional group. While the higher-scoring texts use verb tenses and sentence structures more accurately, the lower-scoring texts contain more frequent problems with verb formation, auxiliaries, agreement and sentence construction.

Table 5. Mechanics Criterion – Traditional Group

Level	**Score Range**	**Participants**	**Percentage**
Excellent to Very Good	5	P2, P20	9.5%
Good to Average	4	P4, P6, P7, P9, P13, P16	28.6%
Fair to Poor	3	P5, P8, P10, P11, P12, P15, P17, P19, P21	42.9%
Very Poor	2	P1, P3, P14, P18	19.0%
Total	-	21	100%
Overall Mean	-	-	3.3 / 5

The results for the mechanics criterion show variation in the use of spelling, accents, capitalization, and punctuation across the traditional writing group. The overall mean is 3.3/5. Two participants (9.5%) are rated "excellent to very good," six (28.6%) "good to average," nine (42.9%) "fair to poor," and four (19.0%) "very poor." Errors involving accents, spelling, capitalization and punctuation occur with different frequencies across the texts. The texts in the "excellent to very good" range (P2, P20) contain very few problems with written conventions. In P2, examples such as *"C'était une journée très agréable"* and *"À la fin du voyage, nous étions fatigués"* show accurate use of accents, apostrophes, capitalization, and punctuation. P20 similarly shows consistent use of these conventions throughout the text. The texts in the "good to average" range (P4, P6, P7, P9, P13, P16) contain only occasional errors in spelling, accents, capitalization or punctuation. P16, for example, uses capitalization and punctuation consistently when presenting place names and different events. Although minor errors occur in some texts, they are relatively limited. The texts in the "fair to poor" range (P5, P8, P10, P11, P12, P15, P17, P19, P21) contain more frequent problems with spelling, accents, or punctuation. P5 includes *"Je commence a pleurer,"* in which the accent required in *"à"* is omitted. Similar errors involving accents, spelling, and punctuation occur in other texts in this range. The texts in the "very poor" range (P1, P3, P14, P18) contain the highest frequency of problems with written conventions. P1 includes forms such as *"je tres content"* and *"vacance tres bien,"* in which the accent in *"très"* is omitted. The text also shows inconsistencies in capitalization and punctuation. Similar problems with written conventions appear in P18. Overall, mechanics is a relatively weak area for the traditional writing group. Most participants fall within the "fair to poor" or "good to average" ranges, while only two are rated "excellent to very good." The most frequent problems involve accents, spelling, capitalization and punctuation.

4.2 Analysis of the ChatGPT-Assisted Writing Group

This section presents the results for the texts produced with ChatGPT support. The texts are evaluated using the ESL Composition Profile developed by Jacobs et al. (1981) across five dimensions: Content, Organization, Vocabulary, Language Use and Mechanics. The findings are presented in tables and illustrated with examples from the students' written productions.

Table 6. Content Criterion – ChatGPT-Assisted Group

Level	Score Range	Participants	Percentage
Excellent to Very Good	30-27	P22, P23, P25, P26, P28, P31, P32, P33, P34, P35, P37, P38, P40, P41	66.7%
Good to Average	26-22	P24, P27, P29, P30, P36, P39, P42	33.3%
Fair to Poor	21-17	-	0%
Very Poor	16-13	-	0%
Total	-	21	100%
Overall Mean	-	-	27.2 / 30

The results for the content criterion indicate a high level of topic development in the ChatGPT-assisted group. The overall mean is 27.2/30, placing the group in the "excellent to very good" range. Fourteen participants (66.7%) are rated "excellent to very good," while seven (33.3%) are rated "good to average." No participant falls within the "fair to poor" or "very poor" categories. Overall, all students address the assigned topic with relevant content and most develop their holiday experiences through specific events, descriptions, and reactions. The texts in the "excellent to very good" range (P22, P23, P25, P26, P28, P31, P32, P33, P34, P35, P37, P38, P40, P41) generally include specific details about activities, settings, feelings and memorable events. P22, for example, develops the experience through *"Nous nous sommes promenés sous les étoiles en écoutant le bruit des vagues"* and *"je me sentais très heureuse, détendue et en paix,"* combining an activity with a description of the participant's feelings. P25 similarly includes *"Nous avons joué au volley sur la plage, pris beaucoup de photos et regardé le coucher du soleil ensemble,"* presenting several activities within the same holiday experience. P31 adds information about cultural activities through *"J'ai découvert plusieurs musées célèbres"* and *"nous avons dégusté des plats locaux dans un petit restaurant."* P33 develops the setting through *"Les rues très animées, des cafés traditionnels, des paysages colorés et des spectacles de rue,"* adding several details about the surroundings. These examples show that the higher-scoring texts develop

the holiday experience by including multiple events, places, activities, descriptions and reactions. Some texts also develop individual events in greater detail. P23 includes *"j'ai commencé à paniquer et j'ai appelé mon frère,"* which describes both the reaction to an event and the action that follows it. P35 similarly includes *"tout le monde essayait de m'aider,"* adding the reaction of other people to the situation. The texts in the "good to average" range (P24, P27, P29, P30, P36, P39, P42) remain relevant to the holiday topic but contain fewer supporting details. P27 includes *"Il a plu plusieurs fois pendant notre séjour,"* adding information about the weather during the trip. P30 states *"Elle aimait beaucoup les spécialités locales et les marchés traditionnels,"* adding details about preferences and local experiences. P36 mentions *"Il y avait beaucoup de bruit dans l'hôtel,"* while P42 refers to an activity through *"J'ai essayé de faire du surf pendant ce voyage."* These details add information about the holiday experience but the surrounding events and descriptions are less extensively developed than in the highest-scoring texts. Overall, content is a strong area in the ChatGPT-assisted group. Most texts develop the holiday experience through several specific events, activities, descriptions and reactions, while the remaining texts contain relevant content with fewer supporting details.

Table 7. Organization Criterion – ChatGPT-Assisted Group

Level	**Score Range**	**Participants**	**Percentage**
Excellent to Very Good	20-18	P22, P23, P24, P25, P26, P28, P29, P30, P31, P32, P33, P34, P35, P37, P38, P39, P40, P41	85.7%
Good to Average	17-14	P27, P36, P42	14.3%
Fair to Poor	13-10	-	0%
Very Poor	9-7	-	0%
Total	-	21	100%
Overall Mean	-	-	18.4 / 20

The results for the organization criterion indicate that the ChatGPT-assisted group generally presents events in a clear and connected sequence. The overall mean is 18.4/20, placing the group in the "excellent to very good" range.

Eighteen participants (85.7%) are rated "excellent to very good," while three (14.3%) are rated "good to average." No participant falls within the "fair to poor" or "very poor" categories. Overall, most students organize the different stages of their holiday narratives in a clear order and use temporal expressions and connectors to link one event to the next. The texts in the "excellent to very good" range (P22, P23, P24, P25, P26, P28, P29, P30, P31, P32, P33, P34, P35, P37, P38, P39, P40, P41) generally show a clear progression from one event to another. P25, for example, uses temporal markers such as *"Au début du voyage…"*, *"Quelques heures plus tard…"*, *"Après le dîner…"*, and *"Avant de rentrer…"* to mark different stages of the narrative. P23 signals a sudden change in the sequence through *"Tout à coup, le bateau a commencé à bouger très fortement,"* introducing a turning point in the event. P38 uses *"Chaque matin…"*, *"En fin d'après-midi…"*, and *"Pendant la soirée…"* to organize activities according to different times of the day. P40 similarly uses *"Après le déjeuner…"*, *"Un peu plus tard…"* and *"À la fin de la journée…"* to connect events in chronological order. P29 uses connectors such as *"Avant notre départ…"*, *"Puis…"*, *"Malgré cela…"*, and *"À la fin…"*, linking events and changes in the narrative more explicitly. The texts in the "good to average" range (P27, P36, P42) also present events in a recognizable order, but use fewer transitions and simpler connections between different parts of the narrative. The sequence can still be followed, although the links between events are less consistently marked than in the highest-scoring texts. Overall, organization is a strong area in the ChatGPT-assisted group. Most texts use temporal expressions and connectors to maintain a clear progression of events, while the lower-scoring texts rely on fewer explicit links between different parts of the narrative.

Table 8. Vocabulary Criterion – ChatGPT-Assisted Group

Level	Score Range	Participants	Percentage
Excellent to Very Good	20-18	P22, P25, P31, P32, P33, P38, P40	33.3%
Good to Average	17-14	P23, P24, P26, P27, P28, P29, P30, P34, P35, P36, P37, P39, P41, P42	66.7%
Fair to Poor	13-10	-	0%
Very Poor	9-7	-	0%
Total	-	21	100%
Overall Mean	-	-	17.0 / 20

The results for the vocabulary criterion show that the ChatGPT-assisted group generally uses enough vocabulary to describe holiday experiences, activities, places and feelings in a clear way. The overall mean is 17.0/20, placing the group in the "good to average" range. Seven participants (33.3%) are rated "excellent to very good," while fourteen (66.7%) are rated "good to average." No participant falls within the "fair to poor" or "very poor" categories. Overall, the texts differ mainly in the range and specificity of the words used for description. The texts in the "excellent to very good" range (P22, P25, P31, P32, P33, P38, P40) show a wider range of descriptive vocabulary. P22, for example, uses *"Le ciel orange du soir"*, *"l'air frais de la montagne"*, and *"une ambiance calme et chaleureuse,"* adding specific details about the setting and atmosphere. P25 includes *"Un souvenir agréable"*, *"des moments amusants"*, and *"un voyage plein d'émotions,"* using different expressions to describe the holiday experience. P31 uses *"Les rues anciennes de la ville"*, *"la culture française"*, and *"les bâtiments célèbres,"* while P33 includes *"camping"*, *"forêt"*, *"animaux sauvages"*, *"cascade"*, and *"activités de plein air."* These choices provide a wider range of words for describing places, activities, surroundings, and experiences. The texts in the "good to average" range (P23, P24, P26, P27, P28, P29, P30, P34, P35, P36, P37, P39, P41, P42) also communicate the intended meaning clearly but use a more limited

range of vocabulary. P24 relies on common expressions such as *"Nous avons mangé"*, *"nous sommes partis"*, and *"nous avons joué."* P37 uses familiar evaluative expressions such as *"Très belles vacances"*, *"temps agréable"*, and *"bons repas."* P36 repeats similar forms in *"Très bons moments"*, *"très contents"*, and *"très amusants,"* while P42 uses basic words such as *"voyage"*, *"famille"*, and *"apprendre à cuisiner."* These expressions communicate the main ideas, but provide less lexical variety than the texts in the highest-scoring range. Overall, vocabulary is a strong area in the ChatGPT-assisted group. Higher-scoring texts use a wider range of specific and descriptive words, whereas the other texts rely more heavily on familiar or repeated expressions.

Table 9. Language Use Criterion – ChatGPT-Assisted Group

Level	Score Range	Participants	Percentage
Excellent to Very Good	25-22	P22, P23, P24, P25, P26, P28, P29, P30, P31, P32, P33, P34, P35, P37, P38, P39, P40, P41	85.7%
Good to Average	21-18	P27, P36, P42	14.3%
Fair to Poor	17-11	-	0%
Very Poor	10-5	-	0%
Total	-	21	100%
Overall Mean	-	-	22.2 / 25

The results for the language use criterion indicate a high level of grammatical accuracy in the ChatGPT-assisted group. The overall mean is 22.2/25, placing the group in the "excellent to very good" range. Eighteen participants (85.7%) are rated "excellent to very good," while three (14.3%) are rated "good to average." No participant falls within the "fair to poor" or "very poor" categories. Overall, most students use verb forms, tense patterns and sentence structures accurately and consistently. The texts in the "excellent to very good" range (P22, P23, P24, P25, P26, P28, P29, P30, P31, P32, P33, P34, P35, P37, P38, P39, P40, P41) generally show accurate tense use and well-formed sentence structures. P22, for example, uses the imparfait appropriately in *"Nous admirions le coucher du soleil"*, *"je me sentais calme et reposée"*, and *"nous*

découvrions plusieurs quartiers anciens." P25 uses a sequence of correctly formed passé composé structures in *"Nous avons quitté la maison très tôt"*, *"nous avons pris le train"*, and *"nous avons installé nos affaires à l'hôtel."* These examples show accurate formation and consistent use of past-tense structures. Some higher-scoring texts also contain more varied sentence structures. P31 includes *"Le lieu que j'ai préféré était un vieux quartier célèbre parce qu'il y avait beaucoup de bâtiments anciens,"* combining a relative clause with a causal clause in the same sentence. P33 uses *"les responsables du camp nous ont montré plusieurs activités,"* while P40 includes *"Plus tard, nous avons discuté avec des touristes étrangers."* These examples show complete sentence structures with correctly formed verbs and appropriate use of complements. The texts in the "good to average" range (P27, P36, P42) also use generally accurate grammatical structures but rely more often on simple sentence patterns. P36 uses *"Nous sommes restés dans un petit hôtel"*, *"Nous jouions sur la plage"* and *"Je passais du temps avec ma famille."* P42 similarly uses *"Ils étaient sympathiques"* and *"J'ai essayé de parler français pendant le voyage."* These sentences use appropriate verb forms and tense patterns, although the range of sentence structures is narrower than in the highest-scoring texts. Overall, language use is a strong area in the ChatGPT-assisted group. Most texts show accurate verb conjugation, appropriate tense use and complete sentence structures, while the lower-scoring texts rely more heavily on simpler sentence patterns.

Table 10. Mechanics Criterion – ChatGPT-Assisted Group

Level	**Score Range**	**Participants**	**Percentage**
Excellent to Very Good	5	P22, P23, P25, P26, P31, P32, P33, P35, P38, P40	47.6%
Good to Average	4	P24, P27, P28, P29, P30, P34, P37, P39, P41	42.9%
Fair to Poor	3	P36, P42	9.5%
Very Poor	2	-	0%
Total	-	21	100%
Overall Mean	-	-	4.4 / 5

The results for the mechanics criterion show that spelling, accents, capitalization, and punctuation are generally used accurately in the ChatGPT-assisted group. The overall mean is 4.4/5. Ten participants (47.6%) are rated "excellent to very good," nine (42.9%) "good to average," and two (9.5%) "fair to poor." No participant falls within the "very poor" category. Overall, errors involving French writing conventions occur relatively infrequently in most texts. The texts in the "excellent to very good" range (P22, P23, P25, P26, P31, P32, P33, P35, P38, P40) show accurate and consistent use of written conventions. P22, for example, includes *"C'était une expérience très agréable"*, *"le soleil brillait toute la journée"*, and *"je me sentais calme et reposée,"* with accurate use of accents and apostrophes. In P25, expressions such as *"Au début du voyage"*, *"Plus tard"*, and *"Comme nous étions très fatigués"* are presented with appropriate punctuation. P31 also includes correctly written expressions such as *"Le parc national"*, *"la vieille bibliothèque"*, *"la place centrale"*, and *"le centre historique."* Similarly, words such as *"spécialité locale"*, *"forêt tropicale"*, *"girafes"*, and *"crocodiles"* are spelled accurately in P32 and P33. The texts in the "good to average" range (P24, P27, P28, P29, P30, P34, P37, P39, P41) also show generally accurate spelling, accentuation, capitalization, and punctuation, although occasional inconsistencies occur. These errors are more frequent than in the highest-scoring texts but do not prevent the texts from being read and understood. The texts in the "fair to poor" range (P36 and P42) contain more noticeable inconsistencies in written conventions. Spelling and punctuation still allow the texts to be understood but errors occur more frequently than in the two higher-scoring categories. Overall, mechanics is a strong area in the ChatGPT-assisted group. Most participants use spelling, accents, capitalization and punctuation accurately, while only two texts show more frequent inconsistencies in these conventions.

4.3 Comparison of the Results of the Two Writing Methods

This section compares the traditional writing group and the ChatGPT-assisted writing group across the five dimensions of the ESL Composition Profile developed by Jacobs et al. (1981): Content, Organization, Vocabulary, Language Use and Mechanics. The comparison provides an overall view of the differences in writing performance between the two groups.

Table 11. Comparison of Mean Scores by Writing Method

Criterion	Traditional Group	ChatGPT-Assisted Group
Content	23.5 / 30	27.2 / 30
Organization	15.4 / 20	18.4 / 20
Vocabulary	14.5 / 20	17.0 / 20
Language Use	17.6 / 25	22.2 / 25
Mechanics	3.3 / 5	4.4 / 5
Total Score	74.3/ 100	89.3 / 100

Note. Total scores were calculated from the unrounded criterion means; therefore, they may differ slightly from the sum of the rounded means presented in the table.

The descriptive results show that the ChatGPT-assisted group has higher mean scores than the traditional writing group across all five dimensions. The overall mean is 89.3/100 for the ChatGPT-assisted group and 74.3/100 for the traditional group. For content, the mean is 23.5/30 in the traditional group and 27.2/30 in the ChatGPT-assisted group. The ChatGPT-assisted texts generally develop the holiday experience through more specific events, descriptions, and reactions, whereas the traditional texts vary more in the amount of supporting detail provided. For organization, the traditional group has a mean of 15.4/20, compared with 18.4/20 for the ChatGPT-assisted group. The ChatGPT-assisted texts more consistently organize events in a clear sequence and use temporal expressions and connectors to link different parts of the narrative. In the traditional texts, events are generally presented in a recognizable order but transitions and explicit links between them are less consistent. The vocabulary mean is 14.5/20 in the traditional group and 17.0/20 in the ChatGPT-assisted group. The ChatGPT-assisted texts generally use a wider range of specific and descriptive words, whereas the traditional texts rely more often on common and repeated vocabulary. For language use, the traditional group has a mean of 17.6/25, while the ChatGPT-assisted group has a mean of 22.2/25. The ChatGPT-assisted texts generally use verb forms, tenses, auxiliaries and

sentence structures more accurately. The traditional texts contain more frequent problems with verb formation, auxiliary use, agreement and sentence construction. For mechanics, the mean is 3.3/5 in the traditional group and 4.4/5 in the ChatGPT-assisted group. Spelling, accents, capitalization, and punctuation are used more consistently in the ChatGPT-assisted texts, whereas errors in these areas occur more frequently in the traditional texts. Overall, the ChatGPT-assisted group has higher mean scores across content, organization, vocabulary, language use and mechanics. These results represent descriptive differences between the two writing conditions and should not be interpreted as evidence that ChatGPT alone causes the higher scores.

5. Discussion

In this study, students' written productions are evaluated across content, organization, vocabulary, language use and mechanics using the ESL Composition Profile developed by Jacobs et al. (1981). The analytic structure of the rubric allows each dimension of writing performance to be considered separately and makes differences between the two writing conditions more visible. Setyowati et al. (2020) indicate that the clear and detailed criteria of the ESL Composition Profile can reduce the margin of error in writing assessment. At the same time, they emphasize that subjectivity arising from the human factor cannot be completely eliminated and that assessors should have experience in educational assessment to support reliable evaluation. In the present study, all compositions are evaluated by the same researcher using the same scoring criteria to maintain consistency throughout the assessment process. The findings show that the traditional writing group generally performs more strongly in content, organization and vocabulary, while greater difficulties appear in language use and mechanics. Problems are particularly evident in verb conjugation, auxiliary use, spelling, accentuation and punctuation. This pattern is broadly consistent with Rufaiqoh et al. (2023), who evaluate Arabic writing using Jacobs' analytic rubric and report stronger performance in content, organization, and vocabulary than in language use and mechanics. Similarly, Aini et al. (2022) report that students' overall writing performance is at the "good to average" level, while difficulties are particularly evident in language use and mechanics. The

grammatical and mechanical problems observed in the traditional writing group therefore resemble those reported in these studies. The ESL Composition Profile also makes it possible to identify strengths and weaknesses across individual dimensions rather than reducing writing performance to a single overall score. Putri (2026) reports that the Jacobs rubric demonstrates a high level of reliability and describes it as an effective instrument for assessing writing performance. In a related context, Turgut and Kayaoğlu (2015) state that the use of rubrics can encourage more active student participation in the writing process and help students recognize the characteristics of effective writing. Although the present study does not examine students' direct use of the rubric during the writing process, the analytic scoring framework provides a detailed basis for identifying strengths and weaknesses across the five dimensions. The comparison between the two groups shows that the ChatGPT-assisted group has higher mean scores in all five dimensions. Compared with the traditional group, the ChatGPT-assisted texts generally show more developed content, clearer sequencing, greater grammatical control, and more consistent spelling and punctuation. These findings represent descriptive differences between the two writing conditions and should not be interpreted as evidence that ChatGPT alone causes the higher scores. The difficulties observed in the traditional writing group also resemble those reported by Patwary et al. (2023) among Bangladeshi students learning English as a foreign language. Their study identifies substantial difficulties in language use, grammar and vocabulary. In the present study, similar limitations are particularly evident in the traditional writing condition, especially in grammatical accuracy, auxiliary use and lexical range. The ChatGPT-assisted group shows stronger performance in these areas, although some variation among students remains. Overall, the findings support the usefulness of an analytic rubric for examining writing performance in detail. Evaluating the five dimensions separately helps identify both group-level differences and dimension-specific strengths and weaknesses that may be less visible in a single overall score.

The findings indicate that ChatGPT-assisted writing may be associated with stronger writing performance across all five dimensions assessed in the study. Compared with the traditional writing group, the ChatGPT-assisted group generally produces more detailed and better-structured texts, follows a clearer progression of events and shows greater lexical variety

and grammatical control. These findings are consistent with Polakova and Ivenz (2024). The researchers indicate that feedback provided by ChatGPT can improve students' writing, particularly in grammatical accuracy, concise expression, the use of foundational knowledge and the accurate use of passive structures. They also emphasize that the rapid and detailed feedback offered by ChatGPT can contribute positively to the writing process. In the present study, the stronger grammatical performance and clearer organization of the ChatGPT-assisted texts show a similar pattern. Mahapatra (2024) likewise reports significant positive effects of ChatGPT on students' academic writing skills. According to the researcher, ChatGPT-assisted feedback contributes to improvements in students' written production, encourages more effective engagement with feedback and helps reduce writing-related anxiety. The study also suggests that ChatGPT can support progress in content development, organization, vocabulary use, and grammatical accuracy. The present findings are broadly consistent with these conclusions, as the ChatGPT-assisted group has higher scores in content, organization, vocabulary, and language use and generally shows more detailed development, more consistent use of transitions and greater grammatical accuracy. Taken together, these similarities suggest that ChatGPT may provide useful support during the writing process by helping students develop ideas, organize their texts, expand lexical choices and improve grammatical accuracy. However, the findings represent differences between the two writing conditions and should not be interpreted as evidence that ChatGPT alone causes the higher performance.

The findings of the present study suggest that the differences associated with ChatGPT-assisted writing extend beyond grammatical accuracy to the organization and structuring of ideas. The ChatGPT-assisted group has higher Organization scores and generally produces texts with clearer sequencing and more consistent use of connectors and transitional expressions. This pattern is consistent with Afra and Jalali (2025) who indicate that ChatGPT-assisted writing instruction is more effective than traditional methods, particularly in content organization, grammar, vocabulary use, coherence and error correction. In the present study, the clearer chronological progression of events, more consistent use of transitional expressions, and more structured development of the ChatGPT-assisted texts reflect a similar pattern. The results also indicate differences in lexical range

between the two groups. The ChatGPT-assisted texts generally contain more varied and descriptive expressions, including *"inoubliable," "magnifique," "expérience incroyable,"* and *"souvenirs précieux."* This finding is consistent with Eid (2025), who reports significant improvements in students' performance through the use of ChatGPT, particularly in vocabulary, grammar, the writing of introductions and conclusions and the accurate application of spelling and punctuation rules. In the present study, the higher vocabulary and mechanics scores of the ChatGPT-assisted group, together with greater lexical variety and more consistent spelling and punctuation, broadly reflect the same pattern. Another relevant pattern is the greater grammatical accuracy observed in the ChatGPT-assisted texts. Since students in this group were permitted to use ChatGPT for grammatical correction and feedback during the writing task, this pattern may be associated with access to corrective feedback during text production. This observation is consistent with Mehrabi and Haghgou (2025) who report that explanatory corrections provided by ChatGPT lead to significant improvements in writing performance, particularly by reducing repeated errors and improving text quality. The researchers also emphasize that artificial intelligence tools should be used consciously and appropriately for educational purposes. The present findings similarly suggest that access to corrective feedback may support greater accuracy in written production, although students' awareness of their own errors is not directly measured in this study. Courant et al. (2024) further indicate that ChatGPT can support students' reading and writing skills in French while also developing their ability to critically analyze and evaluate AI-generated responses. However, critical thinking and evaluative skills are not directly measured in the present study. Therefore, the greater detail and organization observed in the ChatGPT-assisted texts cannot, on their own, be interpreted as evidence of stronger critical thinking or more conscious engagement with the writing process. The present findings are limited to differences in the written products across the dimensions assessed by the ESL Composition Profile.

On the other hand, the findings also point to several considerations regarding the use of ChatGPT in writing. Although the ChatGPT-assisted group generally achieves higher scores, some texts still show limited development or repeated expressions. This observation is consistent with the concerns raised by Mai (2026). The researcher reports that ChatGPT-assisted writing

can improve fluency, grammatical accuracy and organization, while warning that excessive reliance on artificial intelligence may reduce originality and critical thinking. Mai (2026) further emphasizes that students working through traditional methods, although their progress may be slower, may have greater opportunities to develop independent thinking, argumentation skills and a personal writing style. The researcher therefore recommends a balanced use of artificial intelligence together with teacher guidance and activities that encourage critical thinking. Similar concerns are raised by Sain et al. (2025) who note that ChatGPT has limitations, particularly in complex areas requiring specialized expertise. The researchers emphasize that students should not become fully dependent on artificial intelligence tools and should learn to evaluate AI-generated content critically while continuing to develop their own ideas. From this perspective, artificial intelligence is better viewed as a supportive resource that complements rather than replaces human creativity and critical thinking. Lee et al. (2025) likewise report that artificial intelligence tools can support students' writing, particularly in idea generation and content development. At the same time, they stress that such tools should not replace traditional instruction but should be integrated as complementary resources within the teaching and learning process. This perspective is particularly relevant to the present study, in which ChatGPT serves as a support tool rather than as a means of generating complete texts. Overall, the findings show that the ChatGPT-assisted group has higher scores than the traditional group across content, organization, vocabulary, language use and mechanics. The ChatGPT-assisted texts generally show stronger content development, organization, lexical range, grammatical accuracy and control of written conventions. These findings suggest that ChatGPT may provide useful support for written production in French as a foreign language, but they should not be interpreted as evidence that ChatGPT alone causes the higher performance. In line with the literature discussed above, its educational use may be most appropriate when combined with teacher guidance, opportunities for independent writing, and activities that encourage students to evaluate AI-generated suggestions critically.

6. Limitations

The study was conducted with 42 students from a single department and focused on their performance in one writing task. Participants' French proficiency was identified on the basis of the B1-level coursebook used in their regular instruction, without the administration of a separate proficiency test. All compositions were evaluated by the same researcher using the ESL Composition Profile, which was applied consistently to the French written productions without modification of its scoring structure. The analysis focused on descriptive comparisons of the two groups' scores rather than inferential statistical testing. These methodological choices define the scope of the study and should be considered when interpreting the findings.

7. Conclusion and Recommendations

The findings of this study show that the ChatGPT-assisted group achieves higher scores than the traditional writing group across all five dimensions of the ESL Composition Profile. The ChatGPT-assisted texts generally show stronger content development, clearer organization, greater lexical range, higher grammatical accuracy and better control of mechanics. In contrast, the traditional writing group shows greater variation in performance. While some students produce detailed and well-developed narratives, others experience difficulties with idea development, event organization, vocabulary use, grammatical structures, verb conjugation, accentuation and punctuation. At the same time, access to ChatGPT does not result in equally strong performance for all students. Some texts in the ChatGPT-assisted group are less developed or more limited in expression. This variation indicates that the availability of AI support alone does not guarantee a high-quality written product. Since the present study does not directly examine how individual students interact with ChatGPT, differences in performance cannot be attributed to specific patterns of ChatGPT use. Overall, the findings suggest that ChatGPT may serve as a useful support tool for written production in French as a foreign language. However, the higher scores of the ChatGPT-assisted group should be interpreted as differences between the two writing conditions rather than as evidence that ChatGPT alone causes the higher performance. Based on these findings, ChatGPT and similar artificial intelligence tools may be incorporated into French as

a foreign language writing courses as supplementary resources. They can support different stages of the writing process, including idea generation, text planning, vocabulary development, the use of connectors, linguistic correction and feedback. Their use should, however, take place within a structured pedagogical framework and under appropriate teacher guidance. Students should be encouraged to use AI to develop and revise their own texts rather than to generate complete compositions on their behalf. Students should also be encouraged to evaluate AI-generated suggestions critically, verify the accuracy and appropriateness of the information or language provided and remain responsible for their own written production. This approach can help maintain students' active role in the writing process while allowing them to benefit from AI support. The findings also highlight the importance of preparing teachers for the pedagogical use of AI-assisted writing tools. Professional development may help teachers design appropriate AI-supported writing tasks, establish clear guidelines for responsible use and integrate these technologies without replacing students' independent writing practice.

References

Abdelghani, S. (2025). L'exploitation de l'intelligence artificielle pour améliorer la production écrite des futurs enseignants de français à la Faculté de Pédagogie de Tanta (étude expérimentale). *International Journal of Advanced Humanities Research*, *5*(1), 83-115.

Afra, R., & Jalali, V. (2025). An experimental comparison of traditional and ChatGPT assisted writing instruction: Effects on EFL learners' writing performance. *Indonesian EFL Journal: Journal of ELT, Linguistics, and Literature*, *11*(1), 1-17.

Aini, N., Widayanti, D. E., & Winarto, A. E. (2022). An analysis of students' ability in writing cause and effect essay at SMAN 1 Kediri. *Journey: Journal of English Language and Pedagogy*, *5*(1), 64–73.

Bora, P. (2023). Importance of writing skill to develop students' communication skill. *Journal for Research Scholars and Professionals of English Language Teaching*, *7*(35), 1-6.

Brooks, G. (2012). Assessment and academic writing: A look at the use of rubrics in the second language writing classroom. *Kwansei Gakuin University Humanities Review*, (17), 227-240.

Conseil de l'Europe. (2022). *Cadre européen commun de référence pour les langues: apprendre, enseigner, évaluer: Volume complémentaire*. Council of Europe.

Courant, S., Sirichakkaraset, S., & Combet, J. S. (2024). ChatGPT et apprentissage collaboratif: un levier pour la compréhension écrite. *Bulletin de l'ATPF*, *148*(47), 70-92. https://www.tci-thaijo.org/index.php/bulletinatpf/index

Çepni, S. (2011). "Performansların değerlendirilmesi". Emin Karip (Ed.). Performansların değerlendirilmesi (ss.233-285) Ankara: Pegem Akademi.

Çetin, Y. (2011). Reliability of raters for writing assessment: analytic-holistic, analytic-analytic, holistic–holistic. *Mustafa Kemal Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, *8*(16), 471–48

Eid, T. E. M. (2025). Exploitation de l'application de ChatGPT à enrichir la capacité des enseignants non spécialisés à rédiger quelques formes avancées de la production écrite en français. دراسات عربية في التربية وعلم النفس, 156(2), 301-358.

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds & Machines,* 30, 681–694 (2020).

Ghavamian, H., & Dogani, S. (2024). ChatGPT-4: Assistant éducatif innovant pour les enseignants du français. *Recherches en langue française*, *4*(8), 39-58.

Jacobs, H. L., Stephen, A., Zingkgraf, D. R., Wormuth, V., Faye, H., Jane, B., and Hughey. (1981). Testing ESL Composition: A Practical Approach. Rowley: Newbury House Publishers, Inc.

Khalifa, A. S. (2020). Les erreurs de l' Ecrit en classe du FLE. *The Academic Open Journal of Applied and Human Sciences (2709-3344), vol (1), issue (1).*

Kökçü, Y., & İpek, Y. (2025). Yazma becerisinin geliştirilmesinde dijital dönüşüm çağı: Web 2.0 ve yapay zekâ araçları. *Kahramanmaraş Sütçü İmam Üniversitesi Sosyal Bilimler Dergisi*, *22*(2), 705-722.

Lai, W. F., & Fong, M. (2024). Use of comparative research in the study of chemistry education: A systematic analysis of the literature. *Heliyon*, *10*(1).

Laidoudi, A. (2026). ChatGPT dans les mémoires de recherche: étude de la fidélité, la validité et la pertinence des textes générés. *Journal of Languages and Translation*, *6*(1), 318-336.

Lee, Y. J., Davis, R. O., & Choi, J. I. (2025). Integrating generative AI into EFL writing: University students' strategies and perceptions. *Online Journal of Communication and Media Technologies*, *15*(4), e202541.

Mahapatra, S. (2024). Impact of ChatGPT on ESL students' academic writing skills: A mixed methods intervention study. *Smart Learning Environments*, *11*(1), 9.

Mai, H. T. (2026). Evaluating the impact of using ChatGPT on EFL students' argumentative skills. *Discover Education*.

Marshall, C. & Rossman, G. B. (2014). Designing qualitative research. Sage

Mehrabi, M., & Haghgou, Z. (2025). L'impact de ChatGPT sur la production écrite des étudiants iraniens: une analyse d'erreurs. *Didactique du FLES. Recherches et pratiques*, (4: 1), 27-48.

Moulay Omar, A., & Hezabra, H. (2025). Compétence rédactionnelle et Intelligence artificielle: quels enjeux pour quels effets?. *Contextes Didactiques, Linguistiques et Culturels*, *3*(1), 65-87.

Nguyen, T. T. C. (2025). Application de l'intelligence artificielle (ChatGPT) aux méthodes innovantes d'enseignement de la rédaction française à la Faculté de Français de l'Université Phenikaa. In V. V. Dai, B. Marchal, & V. V. Chuong (Eds.), *Le français en Asie-Pacifique : Dynamiques, innovations et perspectives pour l'enseignement, la recherche et la professionnalisation des formations* (pp. 133–146). Éditions de l'Université Nationale du Vietnam à Hanoï.

Patwary, M. N., Reza, M. M., Akhter, H., & Chowdhury, S. (2023). Exploring academic writing proficiency of tertiary EFL learners of Bangladesh: A pedagogical implication. *Integrated Journal for Research in Arts and Humanities*, *3*(5), 89-102.

Polakova, P., & Ivenz, P. (2024). The impact of ChatGPT feedback on the development of EFL students' writing skills. *Cogent Education*, *11*(1), 2410101.

Putri, S. A. W. (2026). the effectiveness of spider diagrams in improving descriptive writing skills of 7th grade students. *RETAIN: Journal of Research in English Language Teaching*, *14*(01), 38-43.

Rufaiqoh, E., Sutaman, S., Arifa, Z., Ibrahim, N. I. E., & Asy'ari, H. (2023). Jacob's analytical assessment in the evaluation of arabic writing skills at Madrasah Aliyah Negeri Program Keagamaan. *Alsinatuna*, *9*(1), 30-40.

Sain, Z. H., Serban, R., Abdullah, N. B., & Thelma, C. C. (2025). Benefits and drawbacks of leveraging ChatGPT to enhance writing skills in secondary education. *At-Tadzkir: Islamic Education Journal*, *4*(1), 40-52.

Setyowati, L., Sukmawan, S., & El-Sulukiyyah, A. A. (2020). Exploring the use of ESL Composition Profile for college writing in the indonesian context. *International Journal of Language Education*, *4*(2), 171-182.

Tate, T. P., Doroudi, S., Ritchie, D., & Xu, Y. (2023). « Educational research and AI-generated writing: Confronting the coming tsunami », pp. 1-20.

Tendero, M., Lenne, A., & Guilloux, J. (2024, April). ChatGPT et troubles dys dans l'enseignement supérieur: Présentation d'un retour d'expériences. In *Diversité, Réussite [s] dans l'Enseignement Supérieur*

Turgut, F., & Kayaoğlu, M. N. (2015). Using rubrics as an instructional tool in EFL writing courses. *Journal of Language and Linguistic Studies*, *11*(1), 47-58.

Algorithmic Literature in the Digital Age: A Discussion on the Literary Competence of Artificial Intelligence

Zeyneb Dolan[1] & *Hakan Dolan*[2]

''The birth of the reader must be at the cost of the death of the Author.''

Roland Barthes

1. Introduction[3]

The act of creating something new has fascinated humanity since the earliest stages of evolution, manifesting itself through the curiosity and pursuit of creation that have shaped human existence (Arathdar, 2021). Owing to its unique and distinctive outcomes, creativity has traditionally been associated with human thought and intuition (Stojanovic et al., 2023). Literary works produced over centuries are likewise regarded as a testament to this human creativity. However, constructing an analogue to the human brain such as artificial intelligence systems capable of generating literary works has taken humans significantly less time than the biological evolution of the human brain itself (Arathdar, 2021). Today, artificial intelligence technologies aim to replicate the human mind by generating contextually responsive outputs, analyzing emotions, and recognizing distinct human characteristics (Shen & Yu, 2021).

[1] Lecturer, Marmara University, School of Foreign Languages, ORCID: 0000-0001-6929-3161, *zeyneb.dolan@marmara.edu.tr*, Türkiye

[2] Lecturer, National Defence University, School of Foreign Languages, ORCID: 0000-0001-9644-2644, *hakanruzgar8@gmail.com*, Türkiye

[3] This study was presented as an oral presentation at the 4th Söylem International Philology Symposium on April 10, 2025.

Reexamining the concept of literature, which has historically been a subject of scholarly debate, has become essential considering the emergence of works generated by generative artificial intelligence. Within approximately a decade, art, literature, music, software, and online content are expected to be predominantly produced by artificial intelligence rather than traditional human authors (Abbott & Rothman, 2023). This digital transformation of literature has opened a new arena of scholarly inquiry centered around the concepts of computational creativity and algorithmic literature. Computational creativity investigates not only the capacity of artificial intelligence to simulate human creativity, but also its potential to produce original, novel, and effective aesthetic artifacts (Baklaga, 2026). While traditional theories of creativity frame literary text production fundamentally as a uniquely human state of intentionality and consciousness, the advent of contemporary generative language models has shifted literature toward an algorithmic axis reshaped through human–machine interaction.

Algorithmic literature extends beyond mere word arrangement to describe an aesthetic production process wherein human literary styles and reflective meanings are successfully emulated through linguistic and statistical patterns (Banerjee, 2026). Recent research clearly demonstrates that literary outputs such as poetry and fiction generated by generative artificial intelligence are indistinguishable from human-authored texts in terms of form, rhyme, style, and reflective language capacity, receiving a comparable degree of reader acceptance. The growing capacity of algorithms to compete robustly with humans in convergent thinking processes places the question of whether literary creativity is an insurmountable boundary unique to the human mind at the center of the sociology of literature (Marrone et al., 2024).

The purpose of this study is to structurally analyze the literary text generation capabilities of generative artificial intelligence within the context of computational creativity and digital humanities, and to examine the theoretical debates surrounding algorithmic literature. Accordingly, the study evaluates AI-generated texts across the genres of poetry, story, and drama based on the criteria of literariness. The ultimate objective of this research is to offer a holistic perspective on emerging paradigms in the sociology of literature by discussing the impact of this technological transformation on

the author's ontological status through the lenses of human–machine co-authorship, originality, ethics, and copyright concerns.

2. The Phenomenon of Authorship in the Age of Artificial Intelligence: Originality, Ethics, and Copyright Concerns

The formal and semantic competence displayed by generative artificial intelligence across various literary genres cannot be evaluated merely as a technological evolution of text production processes. Rather, it signifies a paradigm shift that challenges the sociocultural and philosophical foundations of literary creation. The ability of algorithms to convincingly produce aesthetic forms traditionally attributed to the human mind, such as poetry, short stories, novels, and theatre, complicates the concept of the creative subject behind the literary artifact. Extending beyond the textual quality of the work, this development introduces deeper problematics into the literary world, including how the work should be positioned, who holds ownership over it, and how the creative process itself should be conceptualized.

The rise of algorithmic literature necessitates a redefinition of the author's ontological status, the perception of originality, and the boundaries of literary ownership. As the traditional concept of authorship shifts away from the romanticized figure of the inspired genius toward a facilitator, curator, or co-author interacting with a machine, it generates new points of crisis that must be addressed from both ethical and legal perspectives. The following subsections examine the impact of artificial intelligence on the bond between the author and the text, the implications of human-machine collaboration for literary production, and the debates surrounding copyright, data extraction, and plagiarism emerging in this era of digital creation.

2.1. Authorship and Originality in the Age of Artificial Intelligence

The qualitative and quantitative increase in literary texts produced by artificial intelligence systems has reopened debates surrounding fundamental concepts such as authorial identity, originality, and the ownership of the work. While traditional literary theories regard the author as the creator and sole authority of the text, the integration of large language models into the text production process destabilizes this authority and introduces algorithmic

agency. By raising the question of whether meaning and creativity in literature remain an exclusive privilege of the human mind, this transformation necessitates an ontological redefinition of authorial status (Epstein et al., 2023).

Theoretical objections to authorial authority emerged long before artificial intelligence, originating within poststructuralist thought. Roland Barthes's 1967 essay The Death of the Author dealt a decisive blow to author-centered criticism by arguing that the meaning of a text cannot be reduced to authorial intent or personal biography (Barthes, 1977). According to Barthes, writing is the destruction of every voice and origin, and the author, rather than being a subject embodying internal passions or emotions, functions like an enormous dictionary from which the endless variations of writing are drawn. The operational principle of large language models represents a computational embodiment of Barthes's theoretical arguments. Artificial intelligence is trained on massive datasets comprising the amalgamation of countless texts produced throughout human history, reproducing their structural characteristics, narrative forms, and stylistic conventions through statistical combinations. As discussed by Earp et al. (2024), by automating cultural processes, LLMs demonstrate that the author draws from an already existing network of texts rather than from an innate, internal genius.

While artificial intelligence technically corroborates the "Death of the Author" thesis, it also elevates the reader and the prompter to the position of the ultimate creator of meaning, establishing a new meta-authorship stance (Bajohr, 2024). However, some critics caution that proclaiming the death of the author in the face of artificial intelligence paradoxically reproduces the cult of the author through the machine by creating new technological authorities and idols such as superintelligence. Because the ultimate destination of meaning in Barthes's framework is the reader, any authoritarian notion of an AI Author that monopolizes meaning remains nothing more than a technological replica of the very Author-God that Barthes opposed.

2.2. Human-AI Co-Creativity

The encounter between machine and human intelligence in literature does not necessarily result in an exclusionary rivalry or a relationship of substitution. Current research demonstrates that artificial intelligence can be

positioned not as an autonomous creator destined to replace the author, but as an interactive co-author that supports the literary production process, inspires ideation, and expands narrative boundaries. Human–AI co-creativity can be understood as a hybrid mode of production that combines human creative intention and aesthetic judgment with AI's generative and combinatorial capacities (Boden, 2009; Vinchon et al., 2023). In this process, AI functions not merely as a tool but as a creative partner in ideation and revision, making creative production increasingly iterative and dialogic (Lubart, 2005; Rezwana & Maher, 2023). Hannes Bajohr's (2024) conceptual framework defines this condition in AI systems as "Writing at a Distance." According to Bajohr (2024), in AI-mediated writing processes, the author's intent is not transferred directly to the text itself, but rather into software code, a probabilistic network, and specific prompts. The table below outlines comparable traditional and post humanist paradigms of authorship.

Table 1. A Comparison of Traditional and Post humanist Authorship Paradigms

Conceptual Parameter	Traditional Humanist Authorship	Posthumanist Distributed Authorship (AI)
Source of Creativity	Individual genius, inner inspiration, subconscious processes.	Interaction of statistical probabilities, algorithms, and human guidance.
Agency	Singular, autonomous, and conscious human subject.	Distributed among human, machine, datasets, and platform interfaces.
The Act of Writing	Physical/digital construction of the text, direct writing (inscription).	Instruction of the system, setting rules, guidance (prompts).
The Locus of Meaning	Originates in the mind of the author, is reified in the text.	Meaning is not sought in the machine; it emerges from the relationality between the reader and the algorithmic process.
Originality	Fantasy of creation ex nihilo and uniqueness.	Recombination, intertextuality, synthetic syntheses.

The user reflects their intent through a prompt, which the model transforms through statistical and probabilistic generations, while through the acts of reading, selecting, critiquing, and rewriting, the author evolves into the

reader and editor of their own text. The prompt constitutes a domain of meta-authorship that allows the text to emerge without fully predetermining its conditions in advance. The act of writing has experienced a radical ontological shift from direct inscription to instructing the machine. Bajohr designates this process as the sympoietic author-function. In this framework, the human operates as part of a collective and relational network, initiating the process without serving as the sole arbiter of the outcome. Rather than prioritizing the question of ownership, the ethics of sympoietic authorship demands a focus on process transparency, relational accountability, and interpretive responsibility within distributed creative networks.

2.3. Copyright, Plagiarism, and Ethical Issues in Literary Texts

The integration of generative artificial intelligence into the literary ecosystem introduces profound ethical and legal dilemmas that extend beyond its aesthetic dimensions. The unauthorized use of millions of copyrighted literary works to train artificial intelligence models lies at the center of ongoing debates surrounding data theft and industrial-scale plagiarism. Furthermore, the question of who holds legal ownership over literary texts produced entirely by algorithms demonstrates that existing copyright frameworks struggle to adapt to the speed and nature of algorithmic production (Samuelson, 2023). In the age of artificial intelligence, the author-function faces significant challenges in sustaining its legal and economic roles, such as categorizing texts, attributing accountability, and establishing copyright ownership. This difficulty arises because the system itself can neither possess intentionality nor serve as the recipient of legal liability or reward.

Large language models, which generate coherent texts within seconds in response to user prompts, create an illusion of originality in the human mind, causing ideas distilled from billions of data points to be perceived as personal insight (Lemley & Ouellette, 2025). While destabilizing traditional conceptions of authorship and ownership, this phenomenon rapidly blurs the boundaries between human and machine within creative processes.

In this emerging technological ecosystem, legal scholars and researchers frequently emphasize that the concepts of copyright infringement and plagiarism must not be conflated (Lemley & Ouellette, 2025). Copyright is an

economic and legal entitlement that prevents the unauthorized copying of legally protectable expression. Plagiarism, by contrast, is an ethical violation occurring when another person's ideas, literary expression, or scholarly work are utilized without adequate attribution (Lemley & Ouellette, 2025). Using text generated by artificial intelligence through statistical probabilities may not always constitute legal copyright infringement, as it does not necessarily entail a verbatim duplication of a specific protected work. However, presenting AI-generated ideas or textual passages without transparently acknowledging the machine is regarded directly as plagiarism within academic and literary circles (COPE, 2024).

Concurrently, whether copyrighted literary works can be utilized without authorization in training datasets has escalated into one of the most prominent legal disputes today. In the high-profile lawsuit Authors Guild v. OpenAI (2025), which includes globally recognized authors such as George R.R. Martin and David Baldacci, the plaintiffs contend that OpenAI's unauthorized scanning and ingestion of their works to train language models constitutes large-scale copyright infringement (Sag, 2025; ForensisGroup, 2026). During critical hearings in October 2025, the New York District Court denied OpenAI's motion to dismiss the lawsuit, indicating that an AI system generating summaries that imitate the expressive style of copyrighted fiction could exceed the boundaries of fair use under United States copyright law (Sag, 2025; Madigan, 2025).

From an ethical standpoint, the non-transparent use of generative artificial intelligence creates significant accountability crises in publishing. Although scholars continue to debate whether varying degrees of textual similarity can be tolerated in sections such as methodology, concealing AI-generated ideas from the reader is unequivocally regarded as intentional deception (COPE, 2024). Although language models learn statistical patterns from linguistic form, they should not be assumed to possess human-like semantic understanding and may reproduce historical and social biases embedded in their training data (Bender & Koller, 2020; Caliskan et al., 2017). Such representational biases can contribute to epistemic injustice by systematically disadvantaging particular forms of knowledge, languages, or communities (Helm et al., 2024). In addition, the propensity of AI tools to generate fabricated citations and inaccurate data, commonly referred to

as hallucinations, represents a primary threat to the integrity and reliability of scholarly inquiry (European Commission, 2026).

To address these ethical and legal challenges, universities and academic publishers have begun implementing urgent measures. Leading global publishers such as Elsevier and Wiley have instituted strict policies establishing that artificial intelligence tools cannot be granted authorship status and that any AI software utilized in research must be transparently disclosed within the methodology section (Elsevier). In the emerging digital ecosystem, researchers and writers increasingly require not only technical skills but also AI literacy encompassing the critical and ethical use of AI systems (Long & Magerko, 2020; Pinski & Benlian, 2024). In this context, responsible prompt design can incorporate explicit constraints and verification instructions that help reduce biased or factually unsupported outputs (Dhuliawala et al., 2024; Schick et al., 2021). Whether used in scientific or literary production, AI-generated text should remain subject to critical human oversight, as such systems can produce persuasive content containing fabricated data or references or lacking contextual depth (Gao et al., 2023; Revell et al., 2024; Walters & Wilder, 2023). Human verification and editorial judgment therefore remain essential for preserving the accuracy and intellectual credibility of the resulting text.

3. Criteria of Literariness

This section examines the definition and operational criteria of literariness, tracing its theoretical evolution from Russian Formalism and cognitive reader-response models to Max Bense's Information Aesthetics and recent empirical evaluations of Large Language Models.

3.1. Russian Formalism and the Primacy of Form

Historically, the effort to define what makes a text truly literary begins with Russian Formalism. Rather than analyzing a work through authorial biography, socio-historical contexts, or philosophical messages, the Formalists focused strictly on the linguistic and structural mechanisms of the text itself. Roman Jakobson crystallized this approach in 1921 by arguing that the true subject of literary scholarship is not literature in a broad sense, but literariness (literaturnost) that is, the specific features that distinguish a

literary work from other discourses. Although Russian Formalism encompassed diverse groups and evolving viewpoints, its adherents shared several core axioms (Pilshchikov, 2011, 2015):

- **Autonomy of Verbal Art:** Literature should be examined as an autonomous phenomenon *per se*, rather than as a reflection of external systems such as socioeconomic structures (as in Marxism) or psychological complexes (as in Freudianism).
- **An Independent Discipline:** Literary scholarship should function as an autonomous field governed by systematic and objective methods, akin to the exact sciences.
- **Focus on Form:** Because the unique nature of verbal art lies in its structural organization, studying literature requires the direct investigation of its formal elements and devices.
- **Language as Artistic Material:** Artistic form is created through the intentional processing of raw material. In literature, that material is language itself—worked in the same manner a sculptor shapes clay, wood, or stone. However, underlying narrative structures retain an extra-linguistic dimension and can be retold across different languages and media.
- **Poetic versus Everyday Language:** A fundamental distinction exists between ordinary, practical language and poetic language. Poetic language represents a deliberate and structured deviation from standard everyday communication.
- **Priority of "How" over "What":** The essence of verbal art resides in purposeful formal transformation. Consequently, literary analysis prioritizes questions of construction (*how*) and aesthetic purpose (*for what purpose*) over thematic content (*what*) or external etiology (*why*).

3.2. Cognitive and Empirical Poetics: Miall and Kuiken's Tripartite Model

While classical formalism primarily treated literariness as an intrinsic structural property of the text, modern empirical and cognitive poetics reframes it as an interactive psychological process between the text and the reader. According to Miall and Kuiken (1999), literariness is generated when distinctive stylistic or narrative deviations disrupt a reader's habitual perception of familiar subjects, thereby compelling a cognitive and emotional restructuring of established concepts. To operationalize this construct into an

evaluation framework, their model identifies three interdependent components:

- **Foregrounded Textual Features:** Grounded in Jan Mukařovský's concept of aktualisace, the first component consists of linguistic and structural deviations from normative, everyday discourse. These deviations operate simultaneously across multiple tiers:
  - Micro-stylistic deviations: Phonetic patterns (alliteration, assonance, metric pacing), grammatical irregularities, and semantic shifts (polysemy, unconventional metaphor).
  - Macro-narrative structures: Disrupted chronologies, perspectival ambiguity, sudden shifts in focalization, and free indirect discourse that depart from standard communicative clarity.
- **The Defamiliarizing Psychological Response:** When readers encounter foregrounded structures, automatic and habituated reading patterns are suspended. Attention is arrested, transforming conventional referents into evocative phenomena that resist immediate propositional decoding. Empirical reading-time data demonstrate that segments marked by high foregrounding systematically induce longer processing latencies, heightened uncertainty ratings, and greater perceived strikingness.
- **Affective and Reinterpreted Transformation:** The cognitive pause induced by defamiliarization initiates an affective search after meaning. Readers do not merely process explicit narrative facts; they activate personal, affective schemata to resolve textual ambiguity. Through this process, established existential, ethical, or conceptual frameworks undergo schema refreshment and long-term cognitive modification.

Empirical findings confirm that the cognitive impact of foregrounding remains robust across diverse reader demographics, demonstrating that the perception of literariness is largely independent of formal literary training or specific critical doctrines.

3.3. Computational Foundations: Max Bense's Information Aesthetics

To bridge classical formalist theories with computational text generation, Max Bense's framework of Information Aesthetics introduces mathematical and statistical criteria for evaluating literariness. Developed during the 1950s and 1960s alongside early computer art, Bense integrated semiotics with cybernetics and Shannon's mathematical theory of communication.

Rather than treating literature as an expression of subjective human emotion or authorial genius, Bense redefined the literary work as an objective, measurable informational state. Key principles of Information Aesthetics relevant to algorithmic literature include (Bense, 1960, 1965; Pao, 2023):

- **Language as Combinatorial Material:** Extending the formalist premise of language as raw material, Bense viewed words, letters, and syntax as modular elements arranged via mathematical permutations, probability distributions, and algorithmic rules.
- **Entropy versus Aesthetic Information:** While thermodynamic entropy measures disorder and randomness in physical processes, aesthetic information represents a specific, structured order within linguistic signs. A literary text achieves its aesthetic value through a measurable equilibrium among order, complexity, and redundancy.
- **Transition to Text Theory:** Bense advocated moving from speculative, metaphysical aesthetics to a rational, technical text theory. Within this paradigm, texts are evaluated not through authorial intent, but through the objective statistical distribution of their linguistic signs.

3.4. Large Language Models and Empirical Limits of Computational Creativity

While Information Aesthetics conceptualized literature as an objective arrangement of statistical signs, modern developments in Generative Artificial Intelligence provide the empirical means to test whether probabilistic language generation achieves genuine literariness. Large Language Models operate through autoregressive probability distribution over tokens, producing syntactically fluent and cohesive prose. However, empirical findings by Chakrabarty et al. (2024) demonstrate a substantial disparity between technical fluency and authentic literary creativity.

To evaluate whether machine-generated fiction meets the criteria of literary craft, Chakrabarty et al. (2024) developed the Torrance Test of Creative Writing (TTCW), an artifact-centric evaluation protocol adapted from Torrance's divergent thinking dimensions: Fluency, Flexibility, Originality, and Elaboration. Evaluating 48 short stories written by professional authors (published in The New Yorker) alongside stories generated by leading LLMs (ChatGPT, GPT-4, and Claude v1.3), their expert-administered study revealed several fundamental limitations:

- **The Creativity Gap:** Human-authored stories passed an average of 84.7% of the TTCW creativity tests, whereas LLM-generated narratives passed between 3 and 10 times fewer tests (ranging from 8.7% for GPT-3.5 to 30.0% for Claude v1.3).
- **Absence of Subtext and Rhetorical Complexity:** In line with phenomenological and cognitive criteria of literariness, mature literary discourse relies heavily on implicit meaning, omissions, and subtext. Expert evaluators observed that LLM-generated narratives consistently lack subtext, relying instead on explicit exposition, literal descriptions, and unnatural dialogue that continuously over-explains character emotions.
- **Failure of Defamiliarization:** While formalist theory defines literariness through the intentional defamiliarization of habituated language, LLMs predominantly revert to clichés, overused tropes, predictable plot trajectories, and superficial figurative language that fails to generate genuine aesthetic novelty.
- **Inability to Evaluate Literary Quality:** When LLMs were deployed as evaluators to administer the TTCW benchmark, their scoring exhibited near-zero correlation with human expert assessments, demonstrating an inability to discern the subtle structural and stylistic nuances that constitute literary art.

These empirical findings demonstrate that although algorithmic systems can master statistical syntax and formal surface structures, they struggle to produce the intentional defamiliarization, rich subtext, and cognitive resonance that define authentic literariness.

4. AI and Literary Genres

The application of artificial intelligence across creative literature manifests differently depending on the specific conventions, cognitive demands, and formal architectures of distinct literary genres. While poetry tests computational models against metric constraints, dense figurative language, and interpretative opacity, prose narrative and story creation challenge algorithmic systems to sustain causal logic, long-term plot coherence, and dynamic character development over extended narrative arcs. Meanwhile, the theatrical domain introduces a unique duality that requires algorithmic systems to address both text-based dramatic dialogue and the embodied, real-time dynamics of live stage performance. Examining artificial intelligence across poetry, story writing, and theatre illustrates how computational

models function within diverse aesthetic constraints, highlighting the distinct affordances of human–AI collaboration as well as the enduring psychological, cognitive, and institutional boundaries of machine authorship.

4.1. AI and Poetry

Empirical research on artificial intelligence and poetry generation has focused on how non-expert readers perceive, evaluate, and discriminate machine-generated verse from human-authored literary works. While computational models were historically assumed to lack the capacity for nuanced poetic expression, recent experimental investigations demonstrate that large language models and neural architectures can produce verse that challenges reader discrimination and alters aesthetic evaluation. Specifically, Hitsuwari et al. (2023) examined these dynamics within the strictly constrained 5-7-5 syllable form of Japanese haiku, whereas Porter and Machery (2024) investigated broader English-language poetic traditions spanning ten canonical authors from Geoffrey Chaucer to Dorothea Lasky.

Regarding distinguishability, Porter and Machery (2024) presented 1,634 non-expert readers with poems written by renowned poets alongside imitations generated by ChatGPT-3.5 without human intervention. Participants performed significantly below chance, achieving an overall discrimination accuracy of only 46.6%. Furthermore, Porter and Machery (2024) identified an inverted attribution pattern: readers were statistically more likely to judge AI-generated poems as human-authored than actual human-written poems, with the least recognized human poems being written by canonical authors and the most recognized human poems being generated by AI. In a related investigation, Hitsuwari et al. (2023) tested whether 385 participants could distinguish human-made haiku from those generated by a Long Short-Term Memory (LSTM) system. Their results showed that participants performed at chance level (50% accuracy) when assessing autonomous AI haiku generated without human intervention, but their accuracy dropped significantly below chance to 43% when evaluating AI haiku that had been selected by human evaluators for their beauty.

In terms of aesthetic and qualitative evaluation, both investigations show that machine-generated verse is received favorably by lay audiences. Hitsuwari et al. (2023) compared human-authored haiku against

autonomous AI haiku (human-out-of-the-loop) and human-selected AI haiku (human-in-the-loop). Their analysis revealed that collaborative human-in-the-loop haiku achieved the highest beauty ratings overall, whereas autonomous AI haiku received beauty scores statistically equal to human-made haiku. Additionally, Hitsuwari et al. (2023) found that the perception of beauty in haiku was predicted by psychological dimensions such as imagery, emotional valence, awe, and nostalgia and philosophical dimensions like universality and storytelling. Evaluating fourteen qualitative dimensions across 696 participants, Porter and Machery (2024) found that ChatGPT-3.5 poems were rated significantly higher than canonical human poems on thirteen of the fourteen measures, with the strongest positive effect observed for rhythm and metrical regularity.

The mechanisms driving these evaluations reveal an interaction between reader heuristics and preconceived biases. Porter and Machery (2024) observed a pronounced framing effect: when readers were explicitly told a poem was generated by AI, their qualitative ratings declined significantly compared to when they were told it was human authored, confirming an underlying aversion to machine authorship. Hitsuwari et al. (2023) documented a corresponding manifestation of algorithm aversion, finding a significant negative correlation between the perceived beauty of AI haiku and discrimination accuracy ($r = -0.54$ for autonomous AI haiku and $r = -0.47$ for human-selected AI haiku). This negative relationship indicates that readers operate under the assumption that beautiful works must originate from human creators, leading them to misattribute high-quality AI verse to human poets. To explain this preference, Porter and Machery (2024) argue that language models communicate imagery, mood, and themes in a more direct, accessible, and uncomplicated manner, whereas non-expert readers frequently misinterpret the deliberate opacity and complex metaphors of human poets as algorithmic incoherence. Finally, while Porter and Machery (2024) found that general poetry reading experience did not improve discrimination performance, Hitsuwari et al. (2023) established that specialized haiku experience and individual anthropomorphic tendencies significantly improved participants' ability to identify authors accurately. Hitsuwari et al. (2023) also found that readers who relied on textual consistency were better at identifying human haiku, whereas those

who looked for regularity and intentionality were more successful at identifying AI-generated haiku.

4.2. AI and Story

The integration of artificial intelligence into creative writing has fundamentally reshaped narrative production, shifting the paradigm from solitary authoring to dynamic human–AI co-creation. In a comprehensive systematic review examining the application of AI technologies in story writing between 2018 and 2022, Fang et al. (2023) synthesize the technological mechanisms, pedagogical frameworks, and cognitive outcomes that define this emerging domain. Rather than functioning solely as automated text generators, AI systems in storytelling operate across diverse computational approaches and functional roles, offering both substantial educational benefits and distinct narrative challenges.

4.2.1. Taxonomy of AI Technologies and Functional Roles in Narrative Construction

Fang et al. (2023) categorize AI-supported story writing technologies into three primary functional types: story-writing collaborative agents, storytelling agents, and story animated agents. Story-writing collaborative agents represent most implementations (78%), encompassing tools such as Wordcraft, CoAuthor, BunCho, TaleBrush, and Creative Help. These platforms interact directly with human writers by generating sentence completions, suggesting plot transitions, and expanding narrative settings through open-ended dialogue and natural language interfaces. Storytelling and story-animated agents, by contrast, focus on multimodal and visual narratives, employing tools like SAGA, motion comic storyboards, and 3D game engines such as Unity and Unreal Engine to transform written storylines into interactive, visual, or audio-assisted story experiences.

In terms of functional roles, AI acts predominantly as a story co-creator or collaborator (56%) and an autonomous story generator (30%), with a smaller proportion serving as story animators or storytellers. Under the co-creator paradigm, narrative construction follows an iterative, mixed-initiative process. As illustrated in models like Wordcraft and CoAuthor, the human author establishes the initial premise, setting, or character outlines,

while the AI system generates real-time continuations, assists in planning story arcs, and participates in recursive revision.

4.2.2. Underlying Computational Approaches: Planning-Based versus Machine Learning Models

The architectural mechanisms driving AI story writing bifurcate into two distinct computational traditions: planning-based models and machine learning models.

- **Planning-Based Models:** Planning-based systems utilize goal-directed frameworks which include simulation, global-schema, and multi-agent approaches. These models operate by establishing an initial narrative state and a defined story goal, employing causal reasoners to generate coherent, causal sequences of plot events. This structured approach enables the generation of diverse, logically consistent storylines tailored to specific character motivations.
- **Machine Learning Models:** Accounting for 60% of studies, machine learning approaches rely on statistical probability distributions learned from extensive text corpora. These models encompass script learning and generation, story completion predominantly powered by transformer-based models such as GPT-2, GPT-3, and GPT-Neo, and end-to-end neural story generation using recurrent neural networks and encoder-decoder architectures. These models excel at probabilistic text generation, sentence prediction, and adapting to diverse stylistic conventions based on textual prompts.

4.2.3. Cognitive, Linguistic, and Affective Benefits in Story Creation

The deployment of AI in story writing yields multifaceted benefits across cognitive, linguistic, academic, and affective learning dimensions. Cognitively, collaborative AI agents alleviate common narrative bottlenecks such as writer's block, limited ideation, and difficulty in structuring complex narrative arcs by offering out-of-the-box suggestions for plotlines, character conflicts, and thematic resolutions. Linguistically, AI assistants help non-native and novice writers produce grammatically accurate, lexically

diverse, and structurally coherent prose, reducing cognitive load associated with surface-level mechanics.

On an affective level, Fang et al. (2023) explain the heightened engagement and motivation observed in AI-supported story writing through Keller's ARCS motivational model (Attention, Relevance, Confidence, and Satisfaction). AI writing environments capture learner attention through interactive, responsive interfaces; maintain relevance by aligning narrative generations with the user's personal prompts and background knowledge; build confidence by scaffolding sentence formation when writers encounter creative impasses; and deliver satisfaction through the rapid production of polished, coherent narrative artifacts.

4.3. AI and Theater

The deployment of artificial intelligence in theatre spans both the artistic creation of dramatic works and the operational management of cultural institutions. Investigating text-based dramaturgy, Elias (2026) applied Mark A. Runco's updated creativity framework which incorporates the indices of authenticity and intentionality alongside standard originality and effectiveness to evaluate the creative merit of machine-generated drama. Conducting a comparative stylometric analysis between Jordan Harrison's human-authored play Marjorie Prime and the GPT-2-generated script AI: When a Robot Writes a Play (produced by the THEaiTRE project), Elias (2026) found that while algorithmic text achieved comparable idea density (0.447 versus 0.483), it exhibited significantly lower lexical density, lexical sophistication, and sentence length variation. Consequently, AI-generated dramatic dialogue tends to be literal, oversimplified, and repetitive, lacking the rhetorical subtext and complex character motivations characteristic of human playwriting. Furthermore, because language models depend entirely on pre-existing training datasets, their outputs remain inherently derivative, with intentionality and problem-solving residing strictly in the human operators who design prompts and curate scenes. In live stage productions, this absence of lived consciousness and vocal nuance creates an asymmetric dynamic between robotic actors and spectators, leaving audiences emotionally detached.

From an organizational and practical perspective, Horváth (2025) conducted 24 semi-structured interviews with theatre practitioners, revealing that AI integration occurs predominantly as a bottom-up initiative across operational and support departments. Cultural institutions increasingly deploy machine learning for ticketing data analytics, audience segmentation, personalized performance recommendations, and automated customer service chatbots. In technical and staging domains, AI tools assist in creating visual mood boards, generating dynamic responsive staging, and optimizing lighting systems to improve energy efficiency and environmental sustainability. AI also enhances accessibility through automated real-time subtitling and audio narration for diverse audiences. Within artistic workflows, practitioners utilize AI to support dramaturgical research, summarize historical contexts, and generate preliminary translations, though they emphasize that language models fail to capture the cultural subtleties of human artistic translation and disrupt the spontaneous, fluid exchanges required in live theatrical improvisation.

Despite these operational efficiencies and creative affordances, the expanding presence of AI generates substantial apprehension among theatre professionals regarding labor displacement, digital commodification, and intellectual property theft. Drawing parallels to regulatory protections established during the Hollywood writers' and actors' strikes, practitioners express concern that unmanaged automation risks eroding creative autonomy and devaluing human labor. Consequently, both Elias (2026) and Horváth (2025) conclude that artificial intelligence cannot replace the authentic, culturally grounded voice of human playwrights and performers. To navigate these challenges, Horváth (2025) advocates for transparent disclosure of machine-generated content, interdisciplinary educational initiatives in performing arts programs, and robust regulatory policies that protect copyright while maintaining human creativity at the center of theatrical practice.

5. Conclusion

The integration of large language models and generative algorithms into the text creation process has irrevocably challenged the centuries-old assumption that meaning, agency, and originality in literature and art belong exclusively to humans. This emerging post humanist paradigm marks a

radical shift from the concept of the author as a solitary genius and sole master of the text toward a distributed authorship model guided by algorithmic platforms, where human intent blends with machine statistics.

The criteria of literariness and empirical evaluation tests examined within this study demonstrate that algorithmic systems can successfully replicate structural features at the level of form and syntactic fluency. However, they fall short in constructing rich subtext and achieving defamiliarization, which Russian Formalism and cognitive theories place at the core of literature. The impact of these algorithmic limitations is not uniform across literary genres. While formal constraints in poetry and reader interpretation tendencies statistically blur the boundary between human and machine texts, sustaining causality and character consistency in extended narrative prose, as well as establishing authentic dialogue dynamics in dramatic works, still fundamentally require human structural guidance.

Artificial intelligence can be viewed as a synthetic instrument of plagiarism and pastiche that mechanizes creativity, threatens originality, and reduces linguistic diversity; equally, it can be evaluated as a sympoietic (co-creative) partner that expands the boundaries of literature and transforms linguistic expressions into novel aesthetic forms. While the death of the author philosophically proclaimed by Roland Barthes assumes a statistical and technological reality through LLMs, humanity simultaneously strives to revitalize the author-function in an altered capacity, acting as an instructor, designer, or author at a distance through prompting.

In conclusion, the primary task of literary theory in the age of artificial intelligence is no longer to search for the autonomous human behind the text, but to decipher how symbiotic relationships operate across algorithms, massive datasets, cultural heritage, and human interaction. Ethical frameworks, legal copyright protections, and publishing policies on academic integrity must be updated urgently to accommodate this new distributed agency. Only through such measures can the literary potential of artificial intelligence transition from an existential threat into a hybrid, sustainable intellectual instrument that enhances human creativity.

References

Abbott, R, & Rothman, E. (2023). Disrupting creativity: copyright law in the age of generative artificial intelligence. *Florida Law Review, 75*(6), 1141-1202.

Arathdar, D. (2021). *Literature, narrativity and composition in the age of artificial intelligence*. TRANS-, (27).

Bajohr, H. (2024). Writing at a Distance: Notes on authorship and artificial intelligence. *German Studies Review*, *47*(2), 315–337. https://doi.org/10.1353/gsr.2024.a927862

Baklaga, L. (2026). The impact of AI on creativity: Enhancing human potential or challenging creative expression. *Artificial Intelligence and Applications, 4*(2), 235-247.

Banerjee, R. (2026). Lexical and register configurations of reflective meaning in AI-generated and romantic poetry. *Literariness Journal, 1*(2).

Barthes, R. (1977). The death of the author. In *Image, music, text* (S. Heath, Trans., pp. 142-148). Fontana Press.

Bense, M. (1960). *Programmierung des Schönen: Allgemeine Texttheorie und Textästhetik*. Baden-Baden: Agis.

Bense, M. (1965). *Aesthetica.* Baden-Baden: Agis.

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185–5198). Association for Computational Linguistics.

Boden, M. A. (2009). Computer models of creativity. *AI Magazine, 30*(3), 23–34.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science, 356*(6334), 183–186.

Chakrabarty, T., Laban, P., Agarwal, D., Muresan, S., & Wu, C. (2024). Art or Artifice? Large Language Models and the False Promise of Creativity. In *CHI '24: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (30th ed., pp. 1–34).

Committee on Publication Ethics. (2024). Plagiarism in 2024. *COPE Focus*. Retrieved June 10, 2026, from https://publicationethics.org/topic-discussions/plagiarism-2024

Dhuliawala, S., Komeili, M., Xu, J., Raileanu, R., Li, X., Celikyilmaz, A., & Weston, J. (2024). Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 3563–3578). Association for Computational Linguistics.

Earp, B. D., Yuan, H., Koplin, J., & Mann, S. P. (2025). LLM use in scholarly writing poses a provenance problem. *Nature Machine Intelligence*, *7*(12), 1889–1890.

Elias, S. A. (2026). Algorithmic authorship: evaluating creative merit in human versus AI-generated theatre. *Theatre Research International*, *51*(1), 41–55.

Epstein, Z., Hertzmann, A., Akten, M., Farid, H., Frank, M. R., Groh, M., ... & Mahari, R. (2023). Art and the science of generative AI. *Science, 380*(6650), 1110-1111.

European Commission. (2026). *Living guidelines on the responsible use of generative AI in research*. Retrieved June 10, 2026, from https://research-and-innovation.ec.europa.eu/document/download/2b6cf7e5-36ac-41cb-aab5-0d32050143dc_en

Fang, X., Ng, D. T. K., Leung, J. K. L., & Chu, S. K. W. (2023). A systematic review of artificial intelligence technologies used for story writing. *Education and Information Technologies*, *28*(11), 14361–14397.

ForensisGroup. (2026). Authors Guild et al. v. OpenAI Inc. (2023–present): AI copyright lawsuit over book training data and generative AI. *Expert Legal Witness Blog*. Retrieved June 10, 2026, from https://www.forensis-group.com/resources/expert-legal-witness-blog/authors-guild-et-al-v-open-ai-inc-2023-present-artificial-intelligence-ai-copyright-lawsuit-and-legal-implications-for-creative-works

Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., & Pearson, A. T. (2023). Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. *npj Digital Medicine, 6*, Article 75.

Helm, P., Bella, G., Koch, G., & Giunchiglia, F. (2024). Diversity and language technology: How language modeling bias causes epistemic injustice. *Ethics and Information Technology, 26*, Article 8.

Hitsuwari, J., Ueda, Y., Yun, W., & Nomura, M. (2023). Does human–AI collaboration lead to more creative art? Aesthetic evaluation of human-made and AI-generated haiku poetry. *Computers in Human Behavior*, *139*, 107502.

Horváth, D. (2026). Curtain call for AI: Transforming theatre through technology. *Sustainable Futures*, *9*, 100747.

Lemley, M. A., & Ouellette, L. L. (2025). Plagiarism, copyright, and AI. The University of Chicago Law Review Online. Retrieved June 10, 2026, from https://lawreview.uchicago.edu/online-archive/plagiarism-copyright-and-ai

Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–16). Association for Computing Machinery.

Lubart, T. (2005). How can computers be partners in the creative process: Classification and commentary on the special issue. *International Journal of Human-Computer Studies, 63*(4–5), 365–369.

Madigan, K. (2025). Mid-year review: AI copyright case developments in 2025. *Copyright Alliance*. Retrieved June 10, 2026, from https://copyrightalliance.org/ai-copyright-case-developments-2025

Marrone, R., Cropley, D., & Medeiros, K. (2024). How does narrow AI impact human creativity?. *Creativity Research Journal*, 1-11.

Miall, D. S., & Kuiken, D. (1999). What is literariness? Three components of literary reading. *Discourse Processes*, *28*(2), 121–138.

Pao, L. (2023). Introduction: information aesthetics. *Culture, Theory and Critique/Culture, Theory & Critique*, *64*(4), 367–376.

Pilshchikov, I. (2011). Il retaggio scientifico del formalismo russo e le scienze umane moderne. Trans. by Cinzia Cadamagnani. *Enthymema 5*. 70–102.

Pilshchikov, I. (2015). Nasledie russkoi formal'noi shkoly i sovremennaia filologiia. Antropologiia kul'tury *5*. 319–350

Pinski, M., & Benlian, A. (2024). AI literacy for users – A comprehensive review and future research directions of learning methods, components, and effects. *Computers in Human Behavior: Artificial Humans, 2*(1), 100062.

Porter, B., & Machery, E. (2024). AI-generated poetry is indistinguishable from human-written poetry and is rated more favorably. *Scientific Reports*, *14*(1), 26133.

Revell, T., Yeadon, W., Cahilly-Bretzin, G., Clarke, I., Manning, G., Jones, J., Mulley, C., Pascual, R. J., Bradley, N., Thomas, D., & Leneghan, F. (2024). ChatGPT versus human essayists: An exploration of the impact of

artificial intelligence for authorship and academic integrity in the humanities. *International Journal for Educational Integrity, 20*, Article 18.

Rezwana, J., & Maher, M. L. (2023). Designing creative AI partners with COFI: A framework for modeling interaction in human-AI co-creative systems. *ACM Transactions on Computer-Human Interaction, 30*(5), Article 67.

Sag, M. (2025, November 10). Copyright winter is coming (to Wikipedia?). *Authors Alliance*. Retrieved June 10, 2026, from https://www.authorsalliance.org/2025/11/10/copyright-winter-is-coming-to-wikipedia/

Samuelson, P. (2023). Generative AI meets copyright. *Science, 381*(6654), 158-161.

Schick, T., Udupa, S., & Schütze, H. (2021). Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in NLP. *Transactions of the Association for Computational Linguistics, 9*, 1408–1424.

Shen, Y., & Yu, F. (2021). The influence of artificial intelligence on art design in the digital age. *Scientific Programming*, *2021*, 1–10. https://doi.org/10.1155/2021/4838957

Stojanovic, L., Radojcic, V., Savic, S., Cvetkovic, A. S., & Sarcevic, D. (2023). The influence of artificial intelligence on creative writing: Exploring the synergy between AI and creative authorship. *International Journal of Engineering Inventions, 12*(12), 70–74.

Vinchon, F., et al. (2023). Artificial intelligence & creativity: A manifesto for collaboration. *The Journal of Creative Behavior, 57*(4), 472–484.

Walters, W. H., & Wilder, E. I. (2023). Fabrication and errors in the bibliographic citations generated by ChatGPT. *Scientific Reports, 13*, Article 14045.

About the Series

Contemporary Studies on Digital Humanities, edited by Ali Tilbe, Yusuf Topaloğlu and Gökhan Dinar, is an international, peer-reviewed book series devoted to the study of human expression, knowledge and culture in an age when the tools of reading, writing, teaching and interpreting are increasingly computational. Published by Maurer Press (Frankfurt am Main) in cooperation with AYG Press İstanbul, the series brings together scholars from literary studies, linguistics, language education, translation studies, history, philosophy, cultural and media studies, the arts and the information sciences. It welcomes research on computational approaches to texts and cultures, digital archives and heritage, corpus-based inquiry, artificial intelligence in scholarship, teaching and creative practice, digital pedagogy and assessment, and the ethical and epistemological questions these developments raise. Approaching technology neither as a miracle nor as a threat, but as a new condition of humanistic work, the series seeks to build a plural and rigorous dialogue on the digital transformation of the humanities, with particular attention to research from Türkiye and its neighbouring academic communities alongside contributions from across the globe. It publishes monographs, edited volumes and thematic collections.

About the Editors

Ali Tilbe (Prof. Dr.), Professor of French Language Education and Head of Department at İstanbul University-Cerrahpaşa, Hasan Ali Yücel Faculty of Education. His research focuses on contemporary and Francophone literature, autofiction, sociocriticism and the teaching of French as a foreign language.

Yusuf Topaloğlu (Assoc. Prof. Dr.), Associate Professor of French Language Education at İstanbul University-Cerrahpaşa, Hasan Ali Yücel Faculty of Education. His research focuses on French linguistics and syntax, written production in French as a foreign language and artificial intelligence in language education.

Gökhan Dinar (Asst. Prof. Dr.), Assistant Professor of French Language Education at İstanbul University-Cerrahpaşa, Hasan Ali Yücel Faculty of Education, and the department's Erasmus coordinator. His research focuses on postcolonial Francophone literature, sociocriticism, artificial intelligence in language education and translation studies.

Previously Published

Vol. I: From Generative to Agentic Artificial Intelligence in Foreign Language Teaching, Learning, and Assessment

by Tuncer Can

Vol. II: Between Code and Canon Artificial Intelligence in Language and Literature Studies

by Gökhan Dinar & Yusuf Topaloğlu