RPLTL
Research Papers in Language Teaching and Learning
Hellenic Open University
School of Humanities
Volume 16
March 2026

Generative AI in Language Education: Pedagogical Innovations and Empirical Insights

Guest editors:
Maria Perifanou &
Athanasios Karasimos

PATRAS

Research Papers in *Language Teaching and Learning*

Research Papers in Language Teaching and Learning

Table of Contents of Volume 16, Issue 1, March 2026

EDITORIAL

This special issue of Research Papers in Language Teaching and Learning addresses one of the most timely, dynamic, and consequential developments in contemporary education: the integration of Artificial Intelligence in language education. At a time when AI is rapidly entering educational, academic, and professional domains, often with unprecedented speed and sometimes with insufficient pedagogical reflection, this volume offers a particularly valuable and much-needed scholarly contribution. The papers brought together here do not approach AI as a passing trend or merely as a technical innovation. Rather, they examine it as a phenomenon that is already reshaping the conditions under which languages are taught, learned, assessed, and experienced. Across diverse contexts and perspectives, the contributions in this issue engage with key questions surrounding AI literacy, critical prompting, writing development, feedback cultures, ethical mediation, pedagogical design, and domain-specific applications of large language models. What emerges across the issue is a shared recognition that AI is not merely another technological addition to the language classroom, but a development that compels us to rethink literacy, authorship, feedback, assessment, learner autonomy and agency, teacher mediation, and equity in profound and lasting ways.

The contributions included in this volume reflect the breadth and richness of current scholarship in this area. They examine prompting as a critical literacy practice, the reimagining of feedback cultures in EFL writing, the growing importance of AI literacy in EAP, the creative and critical use of AI in EFL learning, GenAI-driven storytelling and its effects on literacy and writing development, the ethical dimensions of AI in language education, the integration of large language models into corpus-based teaching, and the pedagogical potential of AI-generated Greek dialogues in specialised domains. Taken together, these contributions suggest that the most productive path forward is neither uncritical enthusiasm nor defensive resistance, but informed, reflective, and ethically grounded engagement with the possibilities and limitations of AI in language education. They also demonstrate that the most promising uses of AI are those grounded in human judgment, ethical awareness, reflective practice, and a clear commitment to educational purpose. In this sense, the issue is both highly topical and forward-looking, offering insights of value to researchers, teacher educators, practitioners, and policy stakeholders alike.

I would like at this point to warmly thank the guest editors of this special issue, **Maria Perifanou** and **Athanasios Karasimos**, for their dedicated and highly professional work in bringing this volume to completion. Their editorial leadership, academic commitment, and careful curation of the contributions have been instrumental in bringing together a special issue that is coherent, relevant, and of clear scholarly significance.

At the same time, this issue also marks a moment of transition. As I pass the torch to the next editors, **Vasilios Zorbas** and **Maria Stathopoulou**, I do so with genuine confidence, collegial warmth, and great optimism for the future of *Research Papers in Language Teaching and Learning*. Having worked closely with them for years at the Hellenic Open University, I know their integrity, dedication, and academic seriousness first-hand. I am therefore certain that they will continue to preserve and further enhance the journal's quality, and serve its scholarly community with commitment and care. I warmly wish Vasilis and Maria every success in this important role, and I have no doubt that the journal will

continue to flourish and make a meaningful contribution to research in language teaching and learning.

I would also like to sincerely thank all those who have contributed to the journal during these five years — authors, reviewers, guest editors, and colleagues alike. Their commitment and scholarly contribution have helped define and enrich the volumes published during this period, while also sustaining the quality of RPLTL.

Finally, I would like to acknowledge the founder of the journal, **Nicos Sifakis**, whose vision has shaped RPLTL and guided its course over the years. I am deeply grateful for the confidence he placed in me in appointing me to this role and for his continued support throughout my five years as Editor-in-Chief. These five years have been a rewarding journey, and I feel privileged to have had the opportunity to contribute to the journal's successful course in this role.

Thomai Alexiou

Editor-in-Chief

Research Papers in Language Teaching and Learning
Vol. 16, No. 1, March 2026, 7-12
ISSN: 1792-1244
Available online at http://rpltl.eap.gr

Introduction to the Special Issue on "Generative AI in Language Education: Pedagogical Innovations and Empirical Insights"

Maria Perifanou and Athanasios Karasimos
Guest editors

Generative Artificial Intelligence (GenAI) is rapidly transforming the landscape of language education by opening up new possibilities for personalized instruction, learner engagement, and the automation of selected instructional tasks. With the rise of tools such as ChatGPT, AI-powered chatbots, and adaptive learning platforms, educators are increasingly able to design more dynamic, learner-centred environments that respond to students' needs in real time. These technologies are reshaping how students develop language skills by offering real-time feedback, automated support, authentic conversational practice, and collaborative learning opportunities, while in some cases also reducing anxiety through non-judgmental interaction with AI partners (Crompton et al., 2024; Huang et al., 2022; Kohnke et al., 2023). Among the most promising applications are tools that support speech recognition and pronunciation training, helping learners refine fluency, intonation, and accuracy through immediate, individualized feedback (Belda-Medina & Calvo-Ferrer, 2022). As such tools become more accessible, educators are challenged not only to explore their affordances but also to reflect critically on how AI can be integrated meaningfully into language curricula with attention to inclusion, purpose, and ethics.

GenAI is also playing a growing role in language assessment, with applications that support automated scoring, adaptive testing, and intelligent analysis of learner performance. While these developments may enhance the efficiency and responsiveness of evaluation, they also raise important questions about fairness, transparency, and validity. At the same time, AI-supported learning environments can generate large amounts of learner data and offer new possibilities for timely, individualized feedback and insight into student progress, learning behaviours, and pedagogical effectiveness (Banihashem et al., 2024; Belda-Medina & Calvo-Ferrer, 2022; Crompton et al., 2024; Liu et al., 2023). The potential of AI to boost both formative and summative assessment is considerable, but its implementation requires careful pedagogical judgment, digital literacy among educators, and a strong commitment to learner privacy, autonomy, and trust.

Beyond questions of efficiency, personalization, and assessment, the growing presence of GenAI also invites a broader reconsideration of the relationship between learners, teachers,

and technology. Recent research has moved beyond treating digital tools merely as supportive instruments and now draws on relational perspectives that highlight the distributed and emergent character of agency in language learning environments (Godwin-Jones, 2024; Guerrettaz et al., 2021; Lan & Chen, 2024). From this perspective, learning develops through the interaction of human and technological actors, and AI does not merely assist classroom activity but may also reshape how communication, participation, and knowledge-building take place (Kern, 2018; Lan & Chen, 2024; Godwin-Jones, 2024). Such a shift encourages educators to think more carefully about the pedagogical assumptions embedded in AI use and about the kinds of learning relationships these tools help to create.

From a pedagogical point of view, GenAI can function not only as a feedback mechanism but also as a brainstorming partner, language-learning support, and interactional resource. Recent studies suggest that it can assist learners in drafting, idea generation, revision, and language practice, particularly when its use is guided by clear pedagogical aims and supported by teacher mediation (Boudouaia et al., 2024; Crompton et al., 2024; Guo et al., 2022; Kohnke et al., 2023; Tseng & Warschauer, 2023). It may also contribute to reduced anxiety, greater engagement, and stronger teacher-student rapport in some learning contexts (Ghafouri, 2024; Banihashem et al., 2024). What matters, therefore, is not simply the presence of AI in the language classroom, but the quality of the pedagogical framing that shapes how it is introduced and used.

At the same time, the rapid spread of GenAI makes the development of critical AI literacy increasingly urgent. These tools may produce inaccurate content, reproduce cultural and linguistic bias, and promote standardized language and cultural norms at the expense of more diverse voices, which means that their educational use must be approached with caution, reflection, and a clear sense of social responsibility (Crompton *et al*., 2024; Lan & Chen, 2024). In this evolving landscape, teachers remain central: not only as facilitators of learning, but also as ethical guides who help students engage critically with AI and use it in ways that support meaningful participation, creativity, and human connection (Lan & Chen, 2024). The challenge for language education, then, is not whether AI should be used, but how it can be integrated in ways that remain pedagogically grounded, ethically informed, and responsive to diverse learners and contexts.

This special issue of *Research Papers in Language Teaching and Learning* presents a focused collection of studies united by a timely and urgent theme: the integration of Artificial Intelligence in language education. As AI tools become increasingly embedded in academic and professional contexts, this volume interrogates how they can be adopted critically, ethically, and pedagogically within language teaching and learning. By bringing together theoretical frameworks, empirical research, and classroom-based practice, the contributions collectively map the emerging landscape of AI-mediated language education, its promises, its limitations, and the responsibilities it places on teachers, learners, and institutions.

The volume opens with an investigation into how higher education students in Greece reflected on their use of ChatGPT for critical reading and source evaluation within a digital literacy course. **Panagiota Samioti** draws on thematic analysis of student reflections, forum discussions, collaborative tasks, and questionnaires to identify four interconnected themes: prompting as a metacognitive practice, developing evaluative judgment of AI outputs, recognizing the tool's conditional usefulness, and addressing language-related challenges. Students demonstrated awareness of how prompt formulation influenced response quality and maintained a cautious stance toward AI reliability, highlighting the need for verification and human mediation. The findings position structured and reflective prompting as a means

of fostering metacognitive regulation and critical reading skills, framing ChatGPT not as a substitute for human reasoning but as a catalyst for inquiry and ethical awareness.

The intersection of AI and writing feedback is examined in a theoretical paper by **Ioanna Nifli**, which critiques the dominant "red-pen" feedback paradigm in EFL writing classrooms and advocates for dialogic, co-constructed feedback cultures mediated by both human and algorithmic actors. Drawing on sociocultural theory, feedback literacy, and affect-informed pedagogy, the paper introduces the concept of the Affective-Epistemic Feedback Habitus, a dispositional framework shaped by repeated exposure to authoritative correction and high-stakes evaluation. While LLM-powered tools offer non-judgmental, real-time support capable of scaffolding autonomy and metacognitive engagement, the paper cautions that their use within correction-oriented cultures risks reproducing epistemic dependency. This contribution repositions LLMs as pedagogical interlocutors within culturally responsive, ethically mediated writing instruction.

Questions of AI literacy extend into higher education through an action research study by **Sonia Carmen Munteanu**, which examines how engineering students at a Romanian university engaged with GenAI tools, primarily ChatGPT, for English for Academic Purposes (EAP) homework, in a context without institutional guidance, formal training, or clear curricular provisions. Students predominantly used AI to refine language, clarify ideas, and enhance coherence, reporting positive perceptions of its contribution to efficiency and professional preparedness. However, evidence of more advanced practices such as critical evaluation or iterative prompting remained limited. The study argues for the systematic embedding of AI literacy within EAP curricula to ensure equitable access, ethical use, and pedagogical alignment with an evolving technological landscape.

Complementing these perspectives, **Sophia Kouzouli** presents a values-oriented pedagogical approach to the creative and critical integration of AI in EFL teaching, illustrated through the design and classroom implementation of a lesson entitled "Odyssey 2.0: Values Recharged." Grounded in constructivist, experiential, and inquiry-based learning theories, the lesson design incorporates multimodal activities, digital storytelling, collaborative inquiry, role-play, debate, and creative production, within a learner-centered, values-based educational framework. The paper addresses ethical challenges related to bias, dependency, and data protection, suggesting that meaningful AI integration, when guided by clear pedagogical principles, can enrich EFL learning and contribute to the development of critically aware and socially responsible learners.

The pedagogical value of GenAI-driven storytelling in secondary education is explored by **Sezer Kizilates** through a mixed-methods study conducted in Hong Kong with 76 students randomly assigned to control and experimental groups. While the control group received traditional writing instruction, the experimental group engaged in a five-week GenAI-driven storytelling intervention. The study assessed four dimensions of AI literacy;affective, behavioral, cognitive, and ethical; alongside key writing skills including word usage, narrative structure, creativity, and writing anxiety. Quantitative results demonstrated significant gains in AI literacy and writing performance in the experimental group, alongside reduced writing anxiety. Thematic analysis further revealed heightened motivation, stronger confidence, and greater ethical awareness, positioning GenAI-driven storytelling as a productive pedagogical strategy for developing both language competence and responsible digital literacy in younger learners.

The ethical dimensions of Generative Artificial Intelligence in language education are the focus of a critical theoretical contribution by **Eirini Ioanna Delmadorou**. Drawing on sociocultural theory, Second Language Acquisition, and Constructivism, the paper positions GenAI as a powerful yet non-neutral cultural tool, foregrounding concerns around algorithmic bias, Anglocentric dominance, and digital inequities. In response to these challenges, it proposes the Equitable AI in Language Education Model (EALEM), a conceptual framework built on four guiding principles: inclusivity, transparency, human-in-the-loop mediation, and participatory design. By articulating this framework, the paper contributes to debates on AI in education, providing theoretical and practical guidelines for responsible, equitable, and reflective use of Generative AI in language learning contexts.

A further contribution, by **Athanasios Karasimos, Evangelia-Antonia Efstratiadou, Christos Papatzalas, and Ilias Papathanasiou**, explores the integration of LLMs into corpus-based teaching within specialist programs in Speech-Language Pathology, Computational Linguistics, and Clinical Neurolinguistics. Recognizing the time-intensive nature and technical demands of traditional corpus analysis, particularly with atypical language data such as aphasic speech, the study proposes a systematic educational framework exemplified through the Greek CACLA corpus. Central to this approach is the repositioning of LLMs as "cognitive partners" that handle routine annotation tasks, freeing students to engage in higher-order analysis and critical evaluation of AI outputs. This framework contributes to ongoing efforts to democratize sophisticated linguistic analysis while fostering technological literacy and critical thinking in future practitioners.

The volume closes with a study by **Alexandra Fiotaki**, which examines the use of LLMs to generate Greek banking dialogues for pedagogical purposes. The paper outlines the rationale for selecting banking interactions as a case study and evaluates the output of different LLMs with respect to morphosyntactic accuracy, register appropriateness, pragmatic naturalness, and handling of specialized financial vocabulary. Findings reveal variation across models: some generate fluent but overly generic exchanges, while others handle technical vocabulary well but show inconsistencies in morphological agreement and politeness conventions. Building on these findings, the study proposes morphology-aware dialogue materials for classroom use, including role-plays, error-spotting tasks, and vocabulary development activities, illustrating how AI-generated content can enrich language teaching when subjected to rigorous critical evaluation and purposeful pedagogical design.

Taken together, the contributions to this volume reflect a growing scholarly consensus: the integration of AI in language education is neither straightforward nor inevitable, but requires deliberate pedagogical framing, ethical awareness, and a sustained commitment to learner empowerment. As AI tools continue to evolve and proliferate, this volume offers a timely and rigorous resource for language educators, researchers, teacher trainers, and policymakers seeking to navigate the opportunities and challenges of AI-mediated language learning in an increasingly complex educational landscape.

Acknowledgments

We wish to express our sincere gratitude to the outgoing Editor-in-Chief of Research Papers in Language Teaching and Learning, **Prof. Thomai Alexiou**, for the invitation to guest edit this volume and for her distinguished leadership throughout her five-year tenure. Her academic commitment and careful oversight have been essential to the journal's continued growth and significance in the field. We also join her in acknowledging the journal's founder, **Prof. Nicos Sifakis**, whose vision continues to shape the course of RPLTL.

Furthermore, we are deeply indebted to the following colleagues for their invaluable contribution to the double-blind peer-review process, which ensured the academic rigor of the research presented in this collection:

Anastasios A. Economides, University of Macedonia
Maka Eradze, University of Aquila
Anna Nicolaou, Cyprus University of Technology
Elis Kakoulli Constantinou, Cyprus University of Technology
Panagiotis Kosmas, University of Limassol
Alla Krasulia, Sumy State University
Antigoni Parmaxi, Cyprus University of Technology
Maria Victoria Soule, University of Cyprus
Ana Luísa Mateus Oliveira Chança Torres, Instituto Politécnico de Santarém

Finally, we thank the contributing authors for their insights into the evolving field of AI-mediated language education and extend our warmest wishes to the incoming editors, **Dr. Vasilios Zorbas and Dr. Maria Stathopoulou**. We have every confidence that the journal will continue to flourish under their leadership.

Guest Editors:
Dr. Maria Perifanou, SMILE Lab, University of Macedonia and Aristotle University of Thessaloniki
Dr. Athanasios Karasimos, Aristotle University of Thessaloniki

References

Banihashem, S. K., Kerman, N. T., Noroozi, O., Moon, J., & Drachsler, H. (2024). Feedback sources in essay writing: peer-generated or AI-generated feedback?. *International Journal of Educational Technology in Higher Education, 21*(1), 23. https://doi.org/10.1186/s41239-024-00455-4

Belda-Medina, J., & Calvo-Ferrer, J. R. (2022). Using Chatbots as AI Conversational Partners in Language Learning. *Applied Sciences, 12*(17), 8427. https://doi.org/10.3390/app12178427

Boudouaia, A., Mouas, S., & Kouider, B. (2024). A study on ChatGPT-4 as an innovative approach to enhancing English as a foreign language writing learning. *Journal of Educational Computing Research, 62(*6), 1289-1317. https://doi.org/10.1177/07356331241247465

Crompton, H., Edmett, A., Ichaporia, N., & Burke, D. (2024). AI and English language teaching: Affordances and challenges. *British Journal of Educational Technology, 55,* 2503–2529. https://doi.org/10.1111/bjet.13460

Ghafouri, M. (2024). ChatGPT: The catalyst for teacher-student rapport and grit development in L2 class. *System, 120,* 103209. https://doi.org/10.1016/j.system.2023.103209

Godwin-Jones, R. (2024). Distributed agency in language learning and teaching through generative AI. *Language Learning & Technology, 28*(2), 5–30. https://doi.org/10.64152/10125/73570

Guerrettaz, A. M., Engman, M. M., & Matsumoto, Y. (2021). Empirically defining language learning and teaching materials in use through sociomaterial perspectives. *The Modern Language Journal, 105*(S1), 3-20. https://doi.org/10.1111/modl.12691

Guo, K., Wang, J., & Chu, S. K. W. (2022). Using chatbots to scaffold EFL students' argumentative writing. *Assessing Writing, 54,* 100666. https://doi.org/10.1016/j.asw.2022.100666

Huang, W., 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. https://doi.org/10.1111/jcal.12610

Kern, R. (2018). Five Principles of a Relational Pedagogy: Integrating Social, Individual, and Material Dimensions of Language Use. *Journal of Technology & Chinese Language Teaching, 9*(2). 1–14 .http://www.tclt.us/journal/2018v9n2/kern.pdf

Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for language teaching and learning. *Relc Journal, 54*(2), 537-550. https://doi.org/10.1177/00336882231162868.

Lan, Y. J., & Chen, N. S. (2024). Teachers' agency in the era of LLM and generative AI: Designing pedagogical AI agents. *Educational Technology & Society, 27*(1), I-XVIII. https://doi.org/10.30191/ETS.202401_27

Liu, C., Hou, J., Tu, Y. F., Wang, Y., & Hwang, G. J. (2023). Incorporating a reflective thinking promoting mechanism into artificial intelligence-supported English writing environments. *Interactive Learning Environments, 31*(9), 5614-5632. https://doi.org/10.1080/10494820.2021.2012812

Tseng, W., & Warschauer, M. (2023). AI-writing tools in education: If you can't beat them, join them. *Journal of China computer-assisted language learning, 3*(2), 258-262. https://doi.org/10.1515/jccall-2023-0008

Research Papers in Language Teaching and Learning
Vol. 16, No. 1, March 2026, 13-30
ISSN: 1792-1244
Available online at http://rpltl.eap.gr

Prompting as a Critical Literacy Practice in Higher Education: Student Reflections on Using ChatGPT for Critical Reading and Source Evaluation

Panagiota Samioti

This paper investigates how higher education students reflected on their use of ChatGPT to support critical reading and source evaluation within a higher education digital literacy course in Greece. Drawing on students' written reflections, forum discussions, collaborative tasks and questionnaires, the study explores how students interacted with ChatGPT as a dialogic partner for questioning, verification, and reasoning. Thematic analysis of the students' data identified four interrelated themes: (1) prompting as a metacognitive practice, (2) developing a critical stance through evaluative judgment of ChatGPT's outputs, (3) recognizing the tool's conditional usefulness, and (4) addressing language-related challenges. Findings show that students demonstrated awareness of how the formulation of prompts influenced response quality and depth, using inaccuracies and generalizations as opportunities for deeper analysis. They valued ChatGPT's immediacy and accessibility but maintained a cautious stance toward its reliability, highlighting the need for verification and human mediation. Differences in performance across languages also prompted reflection on linguistic precision and bias. Overall, students' reflections highlight that structured, reflective prompting can foster metacognitive regulation, evaluative judgment, and critical reading, positioning ChatGPT as a catalyst for inquiry and ethical awareness rather than a substitute for human reasoning.

Keywords: ChatGPT; critical reading; source evaluation; metacognition; AI literacy; higher education; language awareness

---

1. Introduction

The rapid emergence of generative artificial intelligence (GenAI) has reshaped the educational landscape, raising urgent questions about pedagogy, ethics, and academic practice. Among these tools, ChatGPT has gained prominence in higher education, where it is used to support feedback, scaffolding, and reflective learning. As its presence expands, understanding how

students critically engage with such technologies becomes essential for developing responsible pedagogical frameworks that enhance human judgment and interpretive skills.

This study addresses this need by examining higher education students' reflections on using ChatGPT (3.5, free version) (OpenAI, 2024/5), as it was available at the time of course implementation within a European-funded higher education project (2022-2025). Although the course centered on academic digital literacy rather than explicit language instruction, it required students to engage in critical reading and source evaluation, practices that are closely tied to language awareness and critical literacy in academic settings. In this study, higher education is treated as a key site of language education, particularly in English for Academic Purposes (EAP) and academic literacy contexts, where critical reading and evaluative judgment are central pedagogical goals. The tasks that supported critical reading and source evaluation were inherently language-based, as they involved interpreting meaning, analyzing discourse, and considering how language choices shape credibility and persuasion in academic communication. Positioning ChatGPT within these language-focused activities allowed students to experience AI not merely as a reading tool but as a medium for reflection, interpretation, and evaluative judgment demonstrating how AI-mediated interaction can help learners question, verify, and interpret meaning across genres and languages in language education contexts.

Specifically, the research examines how students employed ChatGPT to interrogate academic and non-academic texts, identify argumentative and linguistic techniques, and evaluate issues of bias, authorship, and reliability. ChatGPT was not presented as a provider of answers but as a critical interlocutor, that is a tool for prompting inquiry, comparison, and reflection. This framing conceptualizes prompting as both a metacognitive and critical practice through which students cultivate awareness of how language constructs meaning and persuasion. In this study, prompting is understood not as a technical skill but as a critical literacy practice embedded in language-mediated academic activity.

In this study, critical reading and source evaluation are treated as complementary components of critical literacy. While critical reading involves analyzing stance, argumentation, and rhetorical strategy, entailing the ability to identify claims, evaluate reasoning, and interpret perspectives (Flemming, 2012; Barnet & Bedau, 2011, both cited in Sultan *et al.*, 2017) source evaluation focuses on assessing the credibility, bias, and evidential quality of information. Together, they nurture critical linguistic awareness as the ability to interpret and use language reflectively and ethically, which bridges academic literacy and AI literacy.

Drawing on written reflections, forum discussions, collaborative tasks and questionnaires, the analysis identifies four interrelated themes: (1) prompting as a metacognitive practice, (2) critical engagement with AI-generated content, (3) conditional usefulness of GenAI, and (4) language issues in AI outputs. The study therefore aims to explore how structured, reflective use of GenAI, and more specifically ChatGPT's use, can prompt a critical stance enacted through evaluative judgment and verification and support students' development of evaluative judgment, metacognitive regulation, and ethical awareness in higher education. In doing so, it contributes to ongoing debates in applied linguistics and language education about how GenAI can be pedagogically embedded to promote critical reading, source evaluation, and human-centered reasoning in an AI-mediated academic environment.

2. Literature Review

2.1 AI and Academic Literacy

Studies converge on the view that GenAI, such as ChatGPT, can support academic literacy, but its effectiveness depends on the conditions of use. So far, much research has concentrated on writing-related affordances. For instance, in a study, students were enthusiastic in using ChatGPT for essay writing (Šedlbauer, 2024) and, in another, students have appreciated AI support for writing fluency (Yuan et al., 2024) and for content and organization (Mahapatra, 2024; Yuan et al., 2024). Mendez & Tang (2025) report that Swedish students valued its contribution to lower- level outcomes such as summarization and editing. Similarly, positive attitudes have been reported regarding grammar assistance (Johnston et al., 2024; Mahapatra, 2024). However, improvements appear most evident when learners already possess strong critical and ethical foundations, as Baldrich and Domínguez-Oller (2024) note.

Although writing-related affordances dominate literature, research has also explored broader dimensions of academic literacy support, including brainstorming, scaffolding, lecture comprehension, and the development of learner autonomy. Thus, research shows that students expressed appreciation for brainstorming support (Chan & Hu, 2023) and for scaffolding content and knowledge, especially in EFL contexts (Yuan et al., 2024). Undergraduate and postgraduate students have also acknowledged AI's benefits in enhancing lecture comprehension (Kostas et al., 2025), and positive attitudes were reported toward personalized support and language assistance (Chan & Hu, 2023). Also, Mahapatra (2024) underscores ChatGPT's value as a dialogic feedback tool, particularly in contexts where individualized feedback is limited.

Yet, despite these reported benefits, questions of learner autonomy and deeper academic processes remain central. In some studies, learners found the tool less effective for higher-order learning requiring deeper understanding, while other studies highlight risks to the development of critical thinking, academic integrity, and independent problem-solving skills, while warning that overreliance may encourage superficial engagement and academic dishonesty (Mendez & Tang, 2025; Yuan et al., 2024; Kostas et al., 2025; Pitts et al., 2025; Vieriu & Petrea, 2025; Yuan et al., 2024).

Alongside the previous questions, concerns about the quality and reliability of GenAIresponses also persist. These appear in two main forms: some students judge ChatGPT's outputs as overly generic or inaccurate (Šedlbauer, 2024; Chan & Hu, 2023), while others point to cross-lingual inconsistencies, with factual queries answered correctly in some languages but less precisely in others (Xing et al., 2024). These disparities extend to accuracy and timeliness, as updates often appear first in high-resource languages like English. Xing et al. (2024) attribute these differences to uneven training data and limited cross-lingual transfer, noting that stronger translation capabilities correlate with more reliable performance. Consequently, such disparities highlight further pedagogical challenges. On the one hand, working in underrepresented languages may require learners to invest more effort in crafting precise prompts or verifying outputs, potentially enhancing metacognitive and linguistic awareness. On the other hand, inequities risk disadvantaging non-English users. Such inconsistencies complicate students' abilities to summarize texts accurately, evaluate sources, and transfer academic reading or writing strategies across languages

These challenges reinforce the importance of pedagogical mediation and responsible integration of GenAI technologies to scaffold academic literacy. Research stresses that GenAI

should remain embedded in pedagogical frameworks that preserve critical engagement and human interaction. For instance, in English for Academic Purposes settings, Du & Alm (2024) observe that ChatGPT offers flexible opportunities for practice and supports competence, yet they emphasize that teacher-student interaction remains essential for meaningful literacy development. Extending this pedagogical perspective, Anson (2024) argues that large language models (LLMs) can stimulate reflection by generating counterpoints or scaffolds for argument development when positioned as complementary tools. Similarly, another study shows that explicit training in AI use, combined with self- and peer-assessment, helps ensure that ChatGPT supports reflection and deeper engagement with writing rather than serving as a shortcut (Mahapatra, 2024). Finally, at a broader level, recent research highlights the need for institutional policies that promote responsible academic writing practices and uphold pedagogical integrity (Johnston et al., 2024).

2.2 Critical Reading and Source Evaluation with AI

Although extensive research has addressed writing, autonomy, and reliability, fewer studies have examined how AI shapes source evaluation and critical reading, an area to which the present study contributes. In this context, the term "critical reading" is used broadly to encompass students' evaluative and analytical engagement with AI- generated content, including their reasoning, verification, and interpretation of information credibility and relevance.

Within applied linguistics, critical reading is understood as a language-mediated practice through which readers examine how meaning, authority, and interpretation are constructed in texts. Drawing on critical pedagogy, Wallace (2003) conceptualizes critical reading as close attention to linguistic and grammatical choices as resources for reflecting on ideological positioning, credibility, and social context. Extending this view to AI-mediated contexts foregrounds the role of language in evaluating how claims, perspectives, and sources are framed. From a second language pedagogy perspective, critical literacy positions texts as socially situated and open to interrogation, emphasizing learner agency (Luke & Dooley, 2011). Similarly, in EAP, reading is an evaluative practice assessing stance, evidence, and disciplinary credibility (Hyland, 2004).

When viewed through these language-education lens, research on generative AI tools such as ChatGPT indicates that they can scaffold analytical reasoning and source analysis but also risk fostering overreliance and superficial engagement. This duality is evident in studies that frame AI as a collaborative teammate in problem-solving contexts, where students valued its feedback but often overestimated its reliability and cognitive capacity (Marrone et al., 2025). Such misconceptions parallel those observed in critical reading tasks, where uncritical trust in AI-generated information can hinder independent verification and evaluative reasoning.

Nevertheless, when embedded within structured pedagogical strategies, ChatGPT can promote deeper critical engagement. Research shows that EFL learners critically evaluated its responses by posing follow-up questions and verifying information, strengthening understanding and critical thinking (Liang & Wu, 2024). Students also gained confidence in probing questions, analysis, and conceptual understanding, with diverse perspectives challenging assumptions and highlighting the need for refined prompts (Guo & Lee, 2023). Similarly, constructing scientific argument maps around ChatGPT outputs enabled learners to anticipate counterarguments and develop rebuttals, enhancing argumentation skills (Archila et al., 2025). In language education, reflective journals indicate that ChatGPT supported

spatially aware, multi-perspective critical thinking while underscoring the importance of balancing technological support with human reasoning (Liang & Wu, 2024). Finally, students frequently described ChatGPT as a supportive "study partner" offering guidance and feedback (Chan & Hu, 2023; Šedlbauer, 2024).

However, these affordances are constrained by persistent challenges that complicate the picture. For example, research warns that excessive reliance on ChatGPT can discourage students from verifying accuracy or sources, thereby weakening their ability to evaluate authenticity and critical thinking in academic writing, and may even lead to 'cognitive dependence' when outputs are accepted uncritically with little engagement (Yuan et al., 2024; Sari et al., 2025; Suriano et al., 2025). In addition, students may encounter inaccuracies, bias, or fabricated content in AI outputs (Archila et al., 2025; Harrer, 2023). Survey-based studies reinforce these concerns, documenting risks to integrity and critical thinking (Kostas et al., 2025). Finally, many students perceived GenAI outputs as generic (Šedlbauer, 2024), while research warns that efficiency gains with the use of AI technologies may come at the expense of critical depth and authorship clarity (Vieriu & Petrea, 2025; Mendez & Tang, 2025).

These findings reveal a recurring paradox; ChatGPT can foster critical reading and source evaluation skills, when embedded in reflective, teacher-mediated tasks, but risks reinforcing surface-level learning and undermining autonomy when used uncritically. Therefore, effective integration requires explicit attention to AI literacy, institutional guidance, and pedagogical design that encourage students to question, compare, and engage in critical evaluation of outputs rather than passively accepting them (Walter, 2024).

2.3 Prompt Crafting and Metacognitive Awareness

Metacognitive regulation, that is planning, monitoring, and evaluating one's own thinking processes, has long been recognized as essential for effective learning (Flavell, 1987; Schraw & Dennison, 1994, both as cited in Teng, 2024). However, students who depend on AI answers without critical reflection risk falling into "metacognitive laziness" (Sari et al., 2025). This underscores the importance of positioning AI not as an answer-giver but as a dialogic partner, an approach that becomes especially salient in the context of GenAI, where these regulatory processes are most visible through prompt crafting.

Recent studies suggest that effective engagement with ChatGPT depends on learners' reflection on how prompt phrasing influences the relevance and accuracy of responses. Reflective engagement is strengthened when students recognize AI limitations, such as difficulty capturing irony or producing plausible but inaccurate outputs, which often trigger fact-checking, source triangulation, and cautious use (Darwin et al., 2024; Essien et al., 2024; Emran et al., 2024, as cited in Raitskaya & Tikhonova, 2025). Iterative prompting and verification further reinforce monitoring behaviors and metacognitive regulation (Raitskaya & Tikhonova, 2025). Metacognition is therefore central to learners' use of AI feedback (Teng, 2024). Finally, empirical evidence shows that sustained experimentation with prompts promotes higher-level cognitive engagement and deeper critical thinking (Kavadella et al., 2024), especially when guided, as structured mediation can transform superficial AI outputs into opportunities for reflection, elaboration, and higher-order thinking (Borge et al., 2024).

Building on this, research highlights the pedagogical value of structured support in prompting. Guo and Lee (2023) emphasize the importance of crafting clear and specific prompts to obtain more relevant and higher-quality responses. In their study, they implemented a three-stage

ChatGPT-based activity in chemistry courses, that is orientation, essay creation, and output validation, which allowed students to decompose tasks, refine prompts, and verify information. The results showed that detailed instruction was not only crucial for success but also helped students gain confidence in analyzing, evaluating, and drawing logical conclusions. Raitskaya & Tikhonova (2025) also reported that structured interventions around prompt use can cultivate students' critical awareness by requiring them to verify AI-generated responses against other sources. This form of guided practice complements approaches described in their review, such as Lee et al.'s (2024, as cited in Raitskaya & Tikhonova, 2025) guidance-based tool, which used indirect prompts to encourage learners to articulate their reasoning, and Hwang et al.'s (2025, as cited in Raitskaya & Tikhonova, 2025) prompt-based learning model, which strengthened students' ability to generate questions and reflect on their learning processes. Together, these studies underscore that structured interaction with AI, when mediated through carefully designed instruction, enhances both the quality and depth of students' reflective thinking.

Overall, research shows that prompt crafting is a metacognitive practice fostering reflection, regulation, and rhetorical awareness; with structured support and epistemic vigilance it enables deeper learning and critical engagement, but without scaffolding it risks metacognitive laziness, making explicit instruction in prompt formulation and interrogation essential.

3. Methodology

3.1 Research Design

The present study adopted a mixed-methods research design to examine student reflections on the use of GenAI in higher education. The data were produced during the implementation of a hybrid (both in situ and online) six-session academic digital literacy course delivered within a European-funded higher education project (spring semesters 2024 and 2025) focused on critical reading and source evaluation. It is important to emphasize that the course itself was not designed as a research intervention. For the purposes of the present research, the data generated during the course activities were subsequently analyzed to explore how students engaged with ChatGPT.

Specifically, the course aimed to strengthen students' ability to identify the features of credible versus misleading information, distinguish between academic and non- academic or false sources, analyze persuasive and rhetorical language, and reflect on the responsible use of AI tools. Students worked with a variety of materials, including academic journal articles, news media, and social media content, and engaged with digital tools such as the Truly Media Platform (Athens Technology Center & Deutsche Welle., n.d.) and ChatGPT 3.5 (free version) (OpenAI, 2024, 2025). Forty-six undergraduate and postgraduate students from multiple University departments participated. Instruction was in Greek; materials were in Greek and English; responses were in Greek. All were Greek L1 speakers, enabling cross-linguistic reflection on ChatGPT. Despite disciplinary diversity, all developed critical digital literacy through structured engagement with academic and non-academic texts.

3.2 Data Collection and classroom tasks

Four types of data were analyzed:

1. *Written reflections:* Individual or group reflections on students' experiences with ChatGPT during class activities/tasks.
2. *Forum written discussions:* Online debates about the reliability and ethical implications of AI-generated content.
3. *Collaborative written tasks:* Group products from source analysis activities and public-facing texts created to counter misinformation.
4. Quantitative data from *final course evaluation questionnaires*.

Four datasets were analyzed, each derived from the following activities:
• Dataset 1: Group reflections on an activity where students asked ChatGPT which misinformation-related terms (e.g., propaganda, clickbait, manipulation, deepfakes) were most important for students, scientists, and citizens. They compared its responses with their own reasoning and shared their reflections in the forum.
• Dataset 2: Analyses of persuasive strategies in various text genres (e.g., journalistic, multimodal). Students input these texts into ChatGPT to identify rhetorical techniques such as appeals to authority, logic, and emotion.
• Dataset 3: Tasks where students examined contested or fabricated sources across genres. One case involved the academic "chocolate diet" hoax article (Bohannon, 2015), where ChatGPT assessed claim validity and evidence. Other groups explored its evaluations of bias, exaggeration, and fallacies in texts on technology addiction, social media disinformation, and climate news.
• Dataset 4: Evaluations of argumentative writing and the role of language as a persuasive device. ChatGPT was asked to assess whether texts employed emotional vocabulary, hyperbole, or bias, and students compared its outputs with their own judgments.

3.3 Data Analysis

The data were analyzed using thematic analysis (Braun & Clarke, 2006), following a six-phase process of familiarization, coding, theme generation, reviewing, defining, and reporting. Credibility was enhanced through both data-type and methodological triangulation, combining qualitative and quantitative sources to ensure a comprehensive interpretation of findings. Coding was refined iteratively, and thick descriptions of student voices were maintained to preserve authenticity.

An inductive-deductive approach was adopted to capture both emergent and theory-informed patterns in the data. Inductive coding allowed themes to emerge directly from students' reflections and discussions, representing their authentic perspectives. Deductive coding, in turn, was guided by key constructs identified in the literature, such as critical reading and source evaluation, metacognitive awareness and prompting, curiosity and exploratory engagement, autonomy and responsible use, and ethical reflection.

This dual process resulted in four themes: (1) formulating prompts as a reflective and metacognitive practice, (2) developing a critical stance through evaluative judgment of ChatGPT's outputs, (3) recognizing the conditional value of ChatGPT as a learning tool, and (4) addressing language-related challenges and cross-lingual variability in AI performance.

All coding and theme development were conducted by the author, who also taught the course. Although intercoder reliability was not possible, rigor was ensured through prolonged engagement, iterative coding and refinement, systematic data triangulation, and reflexive memoing to address the author's dual instructional and research role.

3.4 Limitations

Several limitations must be acknowledged. First, the academic digital literacy course was a pedagogical initiative within a European-funded project rather than a research intervention. While this enhances ecological validity by capturing authentic classroom practices, it limited design control, as there were no pre- and post-measures, comparison group, or experimental manipulation; therefore, claims about change over time should be treated cautiously. Second, findings derive from a relatively small, context-specific sample from a single Greek higher education course. Although triangulation across reflections, forum discussions, collaborative tasks, and questionnaires strengthens credibility, results are not generalizable beyond this setting. Third, despite data from two cohorts (2024–2025), the study was not longitudinal, and references to increased awareness reflect recurring patterns rather than measured development. Fourth, thematic analysis involves interpretive decisions; transparency and extensive quotations mitigate alternative interpretations. Finally, the author's dual instructor–researcher role may introduce bias, addressed through pattern-focused analysis and verbatim evidence.

3.5 Ethics

The research was conducted in line with institutional and Erasmus+ ethical standards. Students were informed that their course output might be used for research purposes, and informed consent was obtained from all participants. Participation was voluntary, and all data were anonymized before analysis to protect student identity. No identifying details are included in the findings. Since the data were originally generated in a pedagogical setting and later repurposed for research, special care was taken to ensure that consent, confidentiality, and respect for participants' voices were maintained throughout the process. Finally, students were informed about data privacy and terms-of-service considerations when using ChatGPT.

4. Findings

The analysis of written reflections, forum discussions, and collaborative tasks revealed four interrelated themes: (1) formulating prompts, (2) development of critical stance, (3) usefulness with limits, and (4) language issues. Each theme is illustrated through representative quotations that capture key patterns in the data, while only selected student excerpts are presented.

4.1 Prompting as Metacognitive Practice

A strong theme concerns the prompt design process. Students consistently described ChatGPT as useful yet dependent on their own mediation. In a group discussion activity, one group reflected that "with our guidance we can get valuable answers from the ChatGPT concerning arguments" pointing to the need for active involvement. Also, one group highlighted that "the answer changes depending on the role we assign to ChatGPT (e.g., teacher or journalist)", showing how framing influenced responses.

Many students explicitly recognized that the way they phrased their questions shaped the quality of ChatGPT's answers. As one group put it: "The way the question is formulated plays an important role for a clearer and more targeted answer". Another added that "in some cases,

the answers were generalized, while in others more specific depending on the way the question was expressed". Several groups remarked that ChatGPT gave only "general" answers unless their questions were very specific, with some students noting that "the more precise and specific questions we pose, the more personalized and topic-relevant answers it will give", confirming the link between prompt quality and output.

4.2 From Iterative Prompting to Evaluative Judgment: Developing a Critical Stance

A second major theme is the development of critical stance. Students' awareness of the relationship between prompt quality and response depth was reflected in a more deliberate and reflective stance toward their questioning practices. They designed systematic processes to elicit more meaningful and accurate responses from ChatGPT. For instance, when analyzing the "chocolate diet" hoax article (Bohannon, 2015), one student explained:

First, I asked ChatGPT if it knew the techniques of persuasion, and it responded with a specific list. Then I provided the text and asked it to analyze the persuasive techniques. Others described similar step-by-step strategies, such as dividing long texts into smaller parts so that the tool could process them and clarify its understanding: *I had to open a new chat and put in half the article to analyze and then the other* half or posing the task's questions one by one "in the form of questions as they were in the Word document" to obtain more balanced and nuanced answers. Some even highlighted the importance of metacognitive control over the process, noting that "the more precise and specific questions we pose to ChatGPT, the more personalized and topic-relevant answers it will give us". Through such iterative experimentation, students transformed prompt formulation into a deliberate, reflective, and strategic practice rather than a mechanical query-response routine.

Students constantly reflected on the answers that ChatGPT gave. For example, in analyzing the "chocolate diet" hoax article (Bohannon, 2015), they reported that ChatGPT pointed out "logical fallacies… missing details about methodology, no alternative interpretations, and lack of independent confirmation". Students noted that ChatGPT correctly revealed it as a deliberate hoax, which led them to reflect on "how easily scientific articles can be published without strict peer review". As one student concluded, such analyses underscored "the importance of critical thinking and proper evaluation of scientific sources".

Even though students valued ChatGPT's outputs, they positioned themselves as active evaluators rather than passive recipients. Their reflections show that when ChatGPT identified weaknesses in arguments, they treated these not as final answers but as opportunities to sharpen their own critical thinking. For example, students contrasted ChatGPT's answers with their own perspectives, thereby exercising judgment. One student remarked that while the chat claimed an article presented "many logical arguments" their own view was that "the article mentions opinions and research but does not really argue strongly". Another student observed that ChatGPT tended to downplay bias, treating sources as more balanced than they appeared. Finally, one said: "ChatGPT answered that there are a lot of logical arguments, whereas my personal view is that the article mentions opinions and studies but does not argue that much".

4.3 Conditional Usefulness

Regarding the third theme, a group reported that "the answers ChatGPT gave were quite informative and specific" while another highlighted that, even when responses were not fully adequate, "it covered the issue, it just cannot show 100% reliability». Others similarly described the interaction as "quick and clever" but still requiring their own critical thinking.

At the same time, students expressed doubts about ChatGPT's reliability in academic contexts. A recurring issue was inconsistency: "different users received different answers to the exact same question" with responses shifting between generalized statements and more detailed elaborations depending on how prompts were phrased. Also, some students noted misleading or fabricated content. In one case, a group reported that ChatGPT "invented a source that does not exist", while another student was more explicit:*I believe it is very easy to be misled regarding the reliability of a source. Appearances are certainly deceiving.*

As one student concluded: *I think it is a good way at a first stage… but in no case should we trust it, and we must always re-examine the validity of what it states.* Another emphasized that "*regarding reliability, we always need to check sources and not rely only on ChatGPT*". Finally, another added: "*The answer was quite satisfactory, but it requires further investigation*" signaling recognition that ChatGPT could be a useful starting point but shouldn't be considered a unique or definitive source.

Taken together, these qualitative accounts already frame ChatGPT as conditionally useful: valued for speed and initial orientation, but consistently positioned by students as requiring verification, critical judgment, and follow-up inquiry.

This qualitatively articulated cautious stance was also reflected in the final course evaluation questionnaire results (Figure 1), based on the item that asked students: "Is ChatGPT a reliable tool to be used in academic essays?". Responses were almost evenly split: 16 students said "Yes" 14 said "No" and another 16 qualified their answer by saying "It depends on the type of academic essay". Such divided opinions underline the perception of ChatGPT as conditionally useful rather than universally reliable.

The questionnaire item (*Figure 2*) in the same questionnaire that asked students: "To what extent do you trust the information provided by ChatGPT?" reinforces this picture. The vast majority of students (43 out of 46) selected "So and so" indicating a moderate level of trust that depends on context. None reported trusting it "Not at all" while only three students said they "Absolutely" trusted the tool. This distribution confirms that students were willing to engage with ChatGPT but remained skeptical about relying on it without verification.

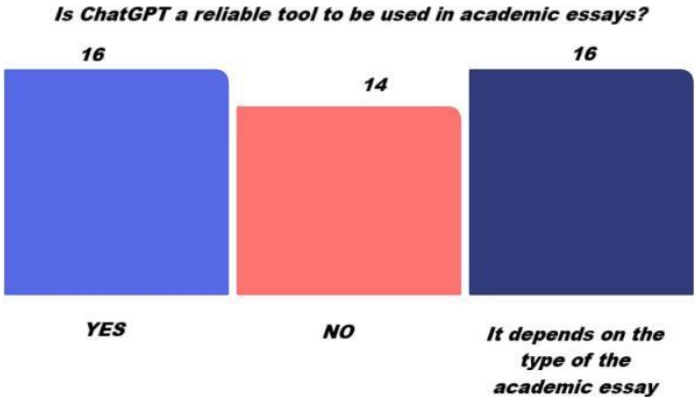


Figure 1: Perceptions of ChatGPT reliability

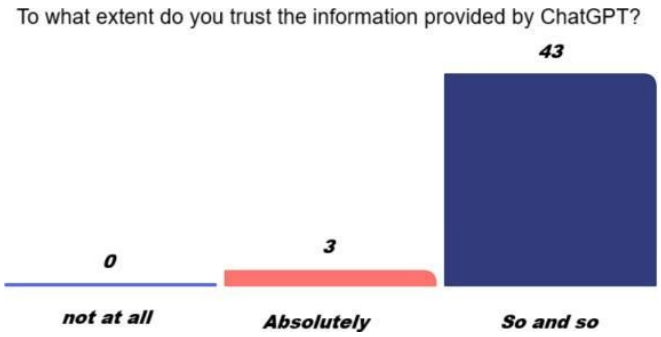


Figure 2: Perceptions of ChatGPT trustworthiness

Taken together, these accounts show that students appreciated ChatGPT's speed, clarity, and perceived usefulness, but they also recognized its limitations in terms of reliability and credibility. Such reflections underline a cautious stance toward integrating AI-generated content into academic tasks, where reliability and critical evaluation remain essential.

4.4 Language-Related Challenges

Several students observed that ChatGPT's performance varied across languages. They noted that responses in English tended to be more complete and more precise, which they attributed to the larger training data available in that language. By contrast, in Greek the tool often struggled, requiring users to phrase questions more carefully in order to obtain meaningful answers.

A group explicitly noted problems with Greek since ChatGPT mistranslated "manipulation" as "transformation" and confused "citizens" with "politicians». Another group observed:
In general, it seems that ChatGPT has difficulty with questions in Greek and it needs the questions to be more specific, or for us to ask more questions.

Finally, some students remarked that, although ChatGPT identified exaggerated or figurative language, "in the linguistic domain it falls short and in no case can we take it into account without critical thinking". Students also reflected on the tone and style of ChatGPT's outputs. Some characterized the responses as "diplomatic, leaving room for further investigation", while others found them overly "robotic", often presented in bullet points or summary form. They noted that "searching in English is more complete and clearer… also less stereotyped and robotic". These observations highlight how language choice and stylistic features influenced their perceptions of ChatGPT's usefulness.

5. Discussion

This study's findings reveal how higher education students used ChatGPT to evaluate sources and cultivate critical reading, highlighting the pedagogical potential of AI- mediated prompting. Viewed through a language-education lens, these practices reflect core concerns of second language reading and critical literacy, including evaluative judgment, metalinguistic awareness, and attention to how language constructs credibility and authority.

Four interrelated topics emerged; prompting as a metacognitive practice, critical engagement, conditional usefulness, and language variability, each revealing how engagement with

ChatGPT can cultivate critical literacy as a co-thinking process rather than a substitute for independent reasoning.

5.1 Prompting as a Metacognitive Practice

Regarding the first topic, students recognized the central role of prompt design and iteration in shaping the specificity and depth of ChatGPT's responses following a pattern which resonates with classic conceptualizations of metacognition; the active planning, monitoring, and evaluation of one's cognitive processes (Flavell, 1987; Schraw & Dennison, 1994, as cited in Teng, 2024). Furthermore, they emphasized that the clarity and framing of their questions determined the quality of the output, reflecting metacognitive awareness of how linguistic choices and question framing shape meaning. This finding aligns with Teng's (2024) argument that metacognitive regulation is integral to AI- supported learning and with Kavadella et al.'s (2024) observation that iterative prompting enhances reflection and higher-order thinking. It also supports Raitskaya & Tikhonova's (2025) view that guided interventions around prompt use promote epistemic vigilance, encouraging learners to monitor and evaluate their interactions critically. Similarly, Lee et al. (2024, as cited in Raitskaya & Tikhonova, 2025) and Hwang et al. (2025, as cited in Raitskaya & Tikhonova, 2025) emphasize that indirect and structured prompts can scaffold students' reasoning processes and deepen reflective engagement. Taken together, these findings demonstrate that deliberate and well-structured prompting nurtures metacognitive control and fosters sustained reflection during AI interaction.

5.2 Critical Engagement and Evaluative Judgment

While the first aspect of metacognition focused on deliberate prompt design, the second was evident when students engaged with ChatGPT's limitations which stimulated additional reflection and evaluation. Similarly to what Darwin et al. (2024), Essien et al. (2024), and Emran et al. (2024) note, the learners in the present study became more reflective when they encountered AI's limitations, turning these shortcomings into opportunities for verification and refinement. More specifically, students reflected on the limitations of ChatGPT's shallow or bulleted outputs and refined them, confirming Borge et al.'s (2024) assertion that ChatGPT can be a simulation tool for higher-order thinking. Therefore, students' active experimentation and reflection counter the risks of "metacognitive laziness" noted by Sari et al. (2025), showing that structured AI use can strengthen both critical reading and metacognitive skills.

Having examined prompting as metacognition, the next theme addresses how critical engagement unfolded through evaluative judgment. Students identified inaccuracies, vague reasoning, and fabricated references in ChatGPT's responses, treating these not as failures but as prompts for deeper analysis. For example, in evaluating the "chocolate diet" hoax article (Bohannon, 2015), they validated ChatGPT's detection of flawed methodology while independently assessing the credibility of the evidence. This engagement supports Liang and Wu's (2024) and Guo & Lee (2023) findings that students can use ChatGPT to stimulate probing questions and strengthen analytical reasoning. It also echoes Archila et al. (2025), whose work demonstrates how engagement with ChatGPT can enhance reasoning and evidence evaluation, showing that these processes can be equally central to critical reading and source assessment. Extending their critical engagement further, students also recognized broader issues of bias and reliability. Their identification of inaccuracy and occasional fabrication in ChatGPT's outputs reflects concerns noted by Harrer (2023), who warns that AI-generated content often reproduces distortions that must be critically interrogated. Students'

recognition of ChatGPT's diplomatic phrasing and tendency to generalize illustrate further their capacity for reflective critique, exemplifying the kind of evaluative stance that Yuan et al. (2024) and Suriano *et al.* (2025) identify as essential to avoid cognitive dependence. These insights are also consistent with Pitts et al. (2025), who caution that uncritical acceptance of AI output may lead to superficial engagement, underscoring the importance of reflective scaffolding. Collectively, these findings confirm Walter's (2024) assertion that AI can foster critical literacy when learners engage in reflective dialogue rather than accept outputs uncritically. Overall, these examples suggest that students used ChatGPT's limitations, that is its occasional superficiality, tendency toward diplomatic answers, and generalizations, as opportunities to interrogate responses, compare them with other evidence, and reflect on standards of reliability and argumentation. In this sense, the tool became a catalyst for sharpening students' evaluative judgment and critical reading skills.

These practices are central to EAP (Hyland, 2004) and second language reading pedagogy, aligning with reading-to-write research that treats locating, evaluating, and using information as core academic practices (Council of Writing Program Administrators, 2014; Grabe & Stoller, 2011; Horning, 2017, as cited in Kocatepe, 2021). From this perspective, students' engagement with ChatGPT reflects established academic reading expectations, where evaluative judgment and source scrutiny are integral to meaning-making and knowledge construction.

5.3 Conditional Usefulness of ChatGPT

The third theme, conditional usefulness, revealed students' nuanced understanding of ChatGPT's strengths and limitations. While they valued its immediacy, clarity, and accessibility, they also emphasized the need for verification and human mediation.

Questionnaire data showed divided trust, with many describing ChatGPT as a useful starting point but not a definitive source. This cautious optimism reflects earlier research reporting both enthusiasm for idea generation and concerns about reliability (Šedlbauer, 2024; Chan & Hu, 2023; Mahapatra, 2024). Following Anson (2024) and Mahapatra (2024), students positioned ChatGPT as a supplementary aid that supports reflection rather than replaces analytical reasoning. At the same time, as Vieriu and Petrea (2025) note, the efficiency gains offered by AI can reduce critical depth and blur authorship boundaries, a limitation that students in this study appeared to address through verification. More specifically, by emphasizing mediation and verification, students demonstrated the form of informed critical judgement envisioned in the course objectives.

5.4 Language Variability and Cross-Linguistic Awareness

The final theme concerned cross-linguistic variability and inequity in ChatGPT's performance. Students observed that English outputs were more complete, precise, and stylistically natural than those produced in Greek, often noting mistranslations, semantic ambiguities, or culturally awkward phrasing in the latter. This observation corroborates findings by Xing et al. (2024) and Kostas et al. (2025), who attribute such disparities to uneven training coverage that privileges high-resource languages like English and limits representational balance across linguistic contexts. Yet rather than viewing these limitations merely as technical flaws, students used them as occasions for critical linguistic reflection and inquiry, adjusting prompts, comparing answers across languages, and refining their phrasing to elicit clearer responses.

These adaptive strategies reflect emerging multilingual AI literacy and heightened metalinguistic awareness, as learners became more conscious of how language mediates meaning, accuracy, and bias in AI interaction. By foregrounding this active negotiation of meaning across languages, the findings extend current understandings of critical literacy in AI-mediated environments and reveal how linguistic variability—though a constraint—can also function as a pedagogical resource. Students' reflections illustrate that such variability encouraged them to interrogate issues of representation, linguistic equity, and the sociocultural dimensions of AI communication, thereby linking language awareness with ethical and critical engagement. These findings align with broader discussions in language education that advocate for ethical, reflective, and pedagogically mediated use of generative tools (Crompton et al., 2024; Du & Alm, 2024; Walter, 2024; Anson, 2024), highlighting that structured classroom tasks and guided mediation play a decisive role in transforming linguistic limitations into opportunities for inquiry and metacognitive growth. When supported by such scaffolding, cross-linguistic challenges become productive sites for reflection on language, bias, and interpretation, reinforcing the view that AI integration in higher education must be designed not only for efficiency but also for linguistic diversity, critical awareness, and ethical learning.

5.5 Implications for practice in language education

ChatGPT as an answer-producing tool, the study highlights the pedagogical value of framing prompting as a language-mediated critical literacy practice embedded in academic reading and evaluation. Prompting is most effective when conceptualized as an iterative process, guiding students from broad exploratory questions to more focused prompts targeting argumentation, evidence, sources, or rhetorical strategies (see Appendix A). Such sequencing foregrounds how prompt formulation shapes analytical depth and supports metacognitive awareness.

AI-mediated prompting is also most effective when integrated into structured critical reading tasks. Comparing ChatGPT's outputs with students' own analyses encourages identification of convergence, omission, and disagreement, positioning ChatGPT as a dialogic resource subject to evaluation and verification. In addition, explicit reflection prompts play a key role in supporting metacognitive regulation by helping students articulate how rephrasing prompts alters responses, what assumptions AI makes, and which claims require external verification.

Finally, cross-linguistic and genre-based prompting fosters critical language awareness by surfacing issues of precision, bias, and contextual appropriateness. Overall, effective integration of ChatGPT in language education depends on conceptualizing prompting as a critical literacy practice that is deliberately designed, scaffolded, and reflected upon, aligning with broader goals of academic literacy and responsible AI use.

6. Conclusion

The data portray students as cautious yet reflective users who balanced appreciation of ChatGPT's affordances with critical awareness of its epistemic and linguistic limitations. These reflective practices were fostered through structured tasks and guided activities that positioned ChatGPT as a dialogic partner rather than an authority. Within this pedagogically mediated framework, students often acted as co- mediators, rephrasing prompts, verifying information, and evaluating the reliability of outputs. As several noted, "with our guidance we

can get valuable answers", acknowledging that effective engagement depended on human judgment and instructional framing. From a language-education perspective, these practices align with pedagogical goals in second language reading and English for Academic Purposes (EAP), where critical evaluation of sources, stance, and evidence constitutes a core outcome of academic literacy rather than a peripheral skill. This pattern supports Du and Alm's (2024) argument that meaningful AI integration relies on pedagogical mediation that sustains purposeful human-AI dialogue and teacher-student interaction. Although direct teacher input was not always visible, it was embedded in the task design that guided students toward reflective inquiry. Such design foregrounded critical reading, evaluative judgment, and learners' agency in meaning-making across languages, illustrating how AI-mediated tasks can be aligned with language-pedagogical priorities rather than efficiency-driven uses of technology. Related research (Johnston et al., 2024) likewise highlights students' need for clear institutional and instructional guidelines, underscoring that pedagogical mediation, whether through teachers or structured learning design, is key to responsible and critical AI use in higher education.

This study shows how generative AI can be integrated in higher education to support critical reading and source evaluation. Students' reflective engagement with ChatGPT fostered metacognitive regulation, evaluative judgment, and critical awareness, positioning higher education, particularly EAP and academic literacy contexts, as key sites of language education and demonstrating GenAI's potential to extend, not replace, human analytical thinking.

For educators and institutions, the implications are clear: effective AI integration requires ethical guidance, attention to linguistic and cognitive diversity, and classroom structures that promote inquiry and verification. For language educators in particular, this entails designing AI-mediated tasks that prioritize critical reading, evaluative judgment, and learners' agency in meaning-making rather than efficiency or automation. Within such frameworks, ChatGPT can act as a catalyst for critical literacy, helping students navigate questions of authorship, credibility, and persuasion in AI-mediated academic communication.

Taken together, the findings demonstrate that structured prompting, reflective evaluation, and cross-linguistic awareness enable students to engage with ChatGPT as a partner in reasoning and reflection. By framing prompting as a reflective, language-mediated practice, the study shows how learners regulate reasoning, interrogate meaning, and evaluate epistemic claims through pedagogical mediation. It also highlights persistent linguistic inequities in AI performance, stressing the need for GenAI literacy that addresses cross-lingual variability and contributes to applied linguistics debates on ethical AI integration in EAP and second language higher education.

Appendix A

Examples of student prompts used with ChatGPT

Task / focus	Initial prompt	Refined / follow-up prompt
Identifying the main claim and evaluating argumentation	Ποιος είναι ο κύριος ισχυρισμός του συγγραφέα; *(What is the author's main claim?)*	Υπάρχουν λογικά σφάλματα στο επιχείρημα του συγγραφέα; Αν ναι, ποια και πώς επηρεάζουν την αξιοπιστία του άρθρου; *(Are there logical fallacies in the author's argument? If so, which ones, and how do they affect the article's credibility?)*

Task / focus	Initial prompt	Refined / follow-up prompt
Analysing persuasive and emotionally loaded language	Το κείμενο χρησιμοποιεί λέξεις υπερβολής; *(Does the text use exaggerated language?)*	Μπορείς να εντοπίσεις συγκεκριμένα παραδείγματα υπερβολής ή συναισθηματικά φορτισμένου λεξιλογίου στο άρθρο και να εξηγήσεις τη λειτουργία τους; *(Can you identify specific examples of exaggeration or emotionally loaded language in the article and explain their function?)*
Evaluating source credibility and evidence	Είναι αξιόπιστη αυτή η μελέτη; *(Is this study reliable?)*	Αξιολόγησε την αξιοπιστία της μελέτης λαμβάνοντας υπόψη το περιοδικό, τους συγγραφείς και τη μεθοδολογία. *(Evaluate the credibility of the study by considering the journal, the authors, and the methodology.)*

References

Anson, D. W. J. (2024). The impact of large language models on university students' literacy development: A dialogue with Lea and Street's academic literacies framework. *Higher Education Research & Development*, *43(7),* 1465-1478. https://doi.org/10.1080/07294360.2024.2332259

Archila, P. A., Molina, J., & Restrepo, S. (2025). Critically reading science-related texts produced by ChatGPT. *Science & Education*, *34(2),* 4627–4662 (2025). https://doi.org/10.1007/s11191-025-00639-y

Athens Technology Center (ATC) & Deutsche Welle. (n.d.). *Truly media*. https://www.truly.media/

Baldrich, K., & Domínguez-Oller, J. C. (2024). The use of ChatGPT in academic writing: A case study in education. Pixel-Bit. *Revista de Medios y Educación*, *71,* 141-157. https://doi.org/10.12795/pixelbit.103527

Bohannon, J. (2015). Chocolate with high cocoa content as a weight-loss accelerator. *Global Journal of Medical Research*, *15(K2),* 9–14. https://medicalresearchjournal.org/index.php/GJMR/article/view/922

Borge, M., Smith, B. K., & Aldemir, T. (2024). Using generative AI as a simulation to support higher-order thinking. *International Journal of Computer-Supported Collaborative Learning*, *19*, 479-532. https://doi.org/10.1007/s11412-024-09437-0

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3(2),* 77-101. https://doi.org/10.1191/1478088706qp063oa

Chan, C. K. Y., & Hu, W. (2023). Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education*, *20(43).* https://doi.org/10.1186/s41239-023-00411-8

Crompton, H., Edmett, A., Ichaporia, N., & Burke, D. (2024). AI and English language teaching: Affordances and challenges. *British Journal of Educational Technology*, *55(6),* 2503-2529. https://bera-journals.onlinelibrary.wiley.com/doi/full/10.1111/bjet.13460

Du, J., & Alm, A. (2024). The impact of ChatGPT on English for Academic Purposes (EAP) students' language learning experience: A self-determination theory perspective. *Education Sciences*, *14(7),* 726. https://doi.org/10.3390/educsci14070726

Guo, Y., & Lee, D. (2023). Leveraging ChatGPT for enhancing critical thinking skills. *Journal of Chemical Education*, *100(12),* 4876–4883. https://doi.org/10.1021/acs.jchemed.3c00505

Harrer, S. (2023). Attention is not all you need: The complicated case of ethically using large language models in healthcare and medicine. *eBioMedicine*, *90,* 104512. https://doi.org/10.1016/j.ebiom.2023.104512

Hyland, K. (2004). *Disciplinary discourses: Social interactions in academic writing*. University of Michigan Press.

Johnston, H., Wells, R. F., Shanks, E. M., Boey, T., & Parsons, B. N. (2024). Student perspectives on the use of generative artificial intelligence technologies in higher education. *International Journal for Educational Integrity*, *20(2),* 1-21. https://doi.org/10.1007/s40979-024-00149-4

Kavadella, A., Dias da Silva, M. A., Kaklamanos, E. G., Stamatopoulos, V., & Giannakopoulos, K. (2024). Evaluation of ChatGPT's real-life implementation in undergraduate dental education: Mixed methods study. *JMIR Medical Education*, *10(1),* e51344. https://doi.org/10.2196/51344

Kocatepe, M. (2021). Reconceptualising the notion of finding information: How undergraduate students construct information as they read-to-write in an academic writing class. *Journal of English for Academic Purposes*, *54*, 101042. https://doi.org/10.1016/j.jeap.2021.101042

Kostas, A., Paraschou, V., Spanos, D., Tzortzoglou, F., & Sofos, A. (2025). AI and ChatGPT in higher education: Greek students' perceived practices, benefits, and challenges. *Education Sciences*, *15(5),* 605. https://doi.org/10.3390/educsci15050605

Liang, W., & Wu, Y. (2024). Exploring the use of ChatGPT to foster EFL learners' critical thinking skills from a post-humanist perspective. *Thinking Skills and Creativity*, *54*, 101645. https://doi.org/10.1016/j.tsc.2024.101645

Luke, A. & Dooley, K. (2011). Critical literacy and second language learning. In E. Hinkel (Ed.), *Handbook of Research in Second Language Teaching and Learning: Volume II,* pp. 856-867,.Routledge.

Mahapatra, S. (2024). Impact of ChatGPT on ESL students' academic writing skills: A mixed methods intervention study. *Smart Learning Environments*, *11*, 9. https://doi.org/10.1186/s40561-024-00295-9

Marrone, R., Zamecnik, A., Joksimovic, S., Johnson, J., & De Laat, M. (2025). Understanding student perceptions of artificial intelligence as a teammate. *Technology, Knowledge and Learning*, *30*, 1847-1869. https://doi.org/10.1007/s10758-024-09780-z

Mendez, J., & Tang, P. (2025). *Student perceptions of achieving intended learning outcomes with generative AI*. [Master's thesis, Lund University]. https://lup.lub.lu.se/luur/download?func=downloadFile&recordOId=9199497&fileOId=9199500

OpenAI. (2024/5). *ChatGPT* (Version 3.5) [Large language model]. https://chat.openai.com/

Pitts, G., Marcus, V. & Motamedi, S. (2025). Student perspectives on the benefits and risks of AI in education. *arXiv preprint.* https://doi.org/10.48550/arXiv.2505.02198

Raitskaya, L., & Tikhonova, E. (2025). Enhancing critical thinking skills in ChatGPT-human interaction: A scoping review. *Journal of Language and Education*, *11(2*), 5-19. https://doi.org/10.17323/jle.2025.27387

Sari, D. N., Marsella, P. E., & Alfiyan, A. R. (2025). Opportunities, challenges and implications of ChatGPT in the self-directed learning process on the critical thinking skills of management students. *International Journal of Integrative Research, 3(7),* 473-488. https://doi.org/10.59890/ijir.v3i7.44

Šedlbauer, J., Činčera, J., Slavík, M., & Hartlová, A. (2024). Students' reflections on their experience with ChatGPT. *Journal of Computer Assisted Learning*, *40(4),* 1526-1534. https://doi.org/10.1111/jcal.12967

Sultan, S., Rofıuddın, A., Nurhadı, N., & Prıyatnı, E. T. (2017).. The effect of the critical literacy

approach on pre-service language teachers' critical reading skills. *Eurasian Journal of Educational Research*, *71*, 159–174.https://izlik.org/JA23GR56JD

Suriano, R., Plebe, A., Acciai, A., & Fabio, R. A. (2025). Student interaction with ChatGPT can promote complex critical thinking skills. *Learning and Instruction*, *95,* 102011. https://doi.org/10.1016/j.learninstruc.2024.102011

Teng, F. M. (2024). Metacognitive awareness and EFL learners' perceptions and experiences in utilising ChatGPT for writing feedback. *European Journal of Education*, *60*, e12811. https://doi.org/10.1111/ejed.12811

Vieriu, A. M., & Petrea, G. (2025). The impact of artificial intelligence (AI) on students' academic development. *Education Sciences*, *15(3),* 343. https://doi.org/10.3390/educsci15030343

Wallace, C. (2003). *Critical reading in language education*. Palgrave Macmillan.

Walter, Y. (2024). Embracing the future of artificial intelligence in the classroom: the relevance of AI literacy, prompt engineering, and critical thinking in modern education. *International Journal of Educational Technology in Higher Education*, *21(15).* https://doi.org/10.1186/s41239-024-00448-3

Xing, X., He, Z., Xu, H., Wang, X., Wang, R., & Hong, Y. (2024). Evaluating knowledge-based cross-lingual inconsistency in large language models. *arXiv preprint*. https://doi.org/10.48550/arXiv.2407.01358

Yuan, Y., Li, H., & Sawaengdist, A. (2024). The impact of ChatGPT on learners in English academic writing: Opportunities and challenges in education. *Language Learning in Higher Education*, *14(1),* 41-56.

Dr Panagiota Samioti (psamioti@uoc.gr) is a linguist and higher education specialist with extensive experience in academic language teaching, language literacy, and multilingual education. She is Laboratory Teaching Staff at the Writing Centre of the University of Crete and a Fellow of the Higher Education Academy (FHEA, UK). She holds a PhD in Theoretical and Applied Linguistics (University of Crete) and an MPhil in English and Applied Linguistics (University of Cambridge). Her expertise spans academic language pedagogy, digital learning environments, and curriculum design in international higher education contexts. She combines a strong research background in linguistics with over twenty-five years of teaching experience across secondary, tertiary, and adult education sectors in Greece and the UK.

Research Papers in Language Teaching and Learning
Vol. 16, No. 1, March 2026, 31-46
ISSN: 1792-1244
Available online at http://rpltl.eap.gr

From Red Ink to Algorithm: Reimagining Feedback Cultures with GenAI in EFL Writing

Ioanna Nifli

Emerging debates in applied linguistics and educational technology have increasingly turned to the transformative role of Generative AI, and more specifically, LLM-powered tools such as ChatGPT, Claude, or Gemini, in reshaping feedback practices in second language writing. Beyond language education, these debates reflect a broader pedagogical reckoning with long-standing behaviourist feedback traditions that prioritise error detection, compliance, and correction over meaning-making and learner agency. Situated within this evolving landscape, the present theoretical inquiry examines how this emerging technology intersects with entrenched feedback cultures in EFL classrooms. The "red-pen" logic of feedback is conceptualised as part of a wider instructional paradigm that frames learning as error elimination through external reinforcement, an orientation historically aligned with behaviourist models of correction, compliance, and measurable accuracy. Drawing on sociocultural theory, feedback literacy, and affect-informed pedagogy, the article critiques the prevailing model of feedback as unidirectional error correction and argues for a shift toward dialogic, co-constructed feedback cultures mediated by both human and algorithmic actors. To account for the deep-rooted patterns that shape learners' orientations toward feedback, the paper introduces the notion of Affective-Epistemic Feedback Habitus—a dispositional framework formed through repeated exposure to authoritative correction, behaviourist assessment logics, and high-stakes evaluation. This habitus is conceptualised as a durable set of emotional and epistemological "defaults" that govern what feedback feels like (e.g., threat, reassurance, shame, safety), what it counts as (e.g., verdict vs. invitation), and how learners are oriented to act upon it (e.g., comply, conceal, negotiate, revise). It shapes how students emotionally and cognitively interpret LLM-mediated feedback, often predisposing them to perceive algorithmic suggestions as final judgments rather than as opportunities for reflection or revision. By exploring the pedagogical affordances and risks of LLM-generated feedback in this context, the paper argues that these tools offer both promise and limitations. While their non-judgmental tone and real-time support can scaffold learner autonomy and metacognitive engagement, their uptake within correction-oriented pedagogical cultures risks reproducing epistemic dependency, instrumental revision practices, and motivational detachment. The article advocates repositioning LLMs not as autonomous evaluators but as pedagogical interlocutors embedded in culturally responsive, ethically mediated writing instruction. Ultimately, it reframes feedback not as a static transmission of correctness, but as a situated, emotionally resonant process co-constructed within specific educational ecologies.

Keywords: Generative AI; Large Language Models; EFL Writing Feedback; Feedback Cultures; Behaviourism; Learner Agency; Feedback Literacy; Dialogic Pedagogy; Sociocultural Theory; Affective Feedback; Teacher Mediation

---

1. Introduction

The rise of Generative Artificial Intelligence (GenAI) in education has reshaped how learners engage with language, writing, and feedback. LLM-powered tools such as ChatGPT, Claude, and Gemini now deliver instant, personalised responses, from grammar correction to idea expansion and stylistic refinement, promising to democratise access to feedback, enhance writing fluency, and support metacognitive growth (Cavaleri et al., 2024:311; Liu & Wang, 2023: 88). Yet this promise is neither neutral nor universally realised. At a systemic level, the emergence of GenAI intersects with a longstanding pedagogical tension across educational contexts between behaviourist models of learning centred on stimulus, correction, and reinforcement, and sociocultural, dialogic approaches that conceptualise learning as meaning-making, participation, and agency.

Historically, the "red pen" has functioned as both a material and symbolic instrument of correction-oriented pedagogy, marking deviation, signalling error, and enforcing compliance (Semke, 1984:195). Across diverse educational systems and subject domains, correction-heavy feedback practices have been shown to undermine learner confidence, narrow revision to surface-level edits, and reduce writing to error avoidance rather than rhetorical development. Research consistently links such practices to heightened anxiety, diminished motivation, and shallow learning outcomes (Horwitz et al., 1986:125; Ryan & Henderson, 2018:884). The red-pen logic thus reflects a broader instructional paradigm in which feedback operates primarily as external regulation rather than as mediated support for learning.

While global debates have increasingly examined the pedagogical and ethical implications of LLM integration, the effects of these technologies remain uneven across educational systems shaped by distinct linguistic, cultural, and assessment traditions. In many EFL contexts—particularly those characterised by high-stakes assessment, strong teacher authority, and accuracy-oriented curricula—feedback practices continue to be dominated by correctional norms and emotionally charged evaluative dynamics. The red-pen model examined in this paper, therefore, reflects a widely recognisable educational habit of positioning feedback as behavioural control: detecting deviation, marking error, and rewarding conformity. When feedback functions primarily as external regulation, learner attention is systematically narrowed to error avoidance rather than to the development of voice, meaning, and rhetorical agency.

In exam-oriented EFL systems, writing pedagogy has frequently prioritised grammatical accuracy and lexical appropriateness over rhetorical development or authorial voice (Lyrigkou, 2021). Feedback in such contexts is often delivered through visibly corrective, hierarchical, and emotionally consequential practices that many learners experience as judgment rather than an invitation to revise, fostering anxiety, demotivation, and superficial revision strategies. Within these instructional ecologies, LLM-generated feedback holds the potential both to disrupt established norms and to amplify them.

Although GenAI tools can prompt more frequent revision and quicker feedback, their pedagogical impact is not inherent but contingent on how they are framed, mediated, and taken up. In correction-oriented

feedback cultures, learners may default to using GenAI as an automated grammar checker, while teachers may tacitly legitimise this use by aligning AI feedback with error elimination rather than metacognitive engagement. Under these conditions, GenAI risks functioning as a "digital red pen": efficient, immediate, and scalable, yet pedagogically reductive. This risk does not arise because GenAI is intrinsically correctional, but because existing feedback scripts, shaped by behaviourist legacies, assessment incentives, and authority-driven norms, can domesticate a dialogic technology into an instrumental workflow. In high-stakes, exam-driven environments, the path of least resistance is often instrumental use rather than dialogic meaning-making.

This paper, therefore, treats LLM adoption as a grand challenge of feedback culture design across EFL contexts: whether new technologies will be used to extend corrective regimes, accelerating error detection and compliance, or to cultivate dialogic, learner-centred revision practices that redistribute interpretive agency.

Accordingly, the paper critically examines how LLM-powered tools interact with feedback cultures that is, the normative and affective patterns shaping how feedback is given, received, and interpreted in localised pedagogical settings (Carless & Winstone, 2020:153). Drawing on sociocultural theory, feedback literacy, and affect-informed pedagogy, it argues that feedback is not merely technical text improvement but a relational, culturally embedded act reflecting broader pedagogical ideologies. Particular attention is paid to how behaviourist feedback traditions may be inadvertently reproduced through GenAI when critical AI literacies and teacher mediation are absent. The analysis explores how LLMs may either reinforce hierarchical, judgmental paradigms or foster dialogic, agency-oriented models of feedback, positioning them not as autonomous correctors but as scaffolds for learner authorship, metacognition, and emotional safety when mediated with pedagogical and cultural sensitivity.

2. Theoretical Framework: Feedback Cultures and Sociocultural Perspectives

2.1 Feedback Cultures

In recent years, the concept of feedback cultures has emerged as a critical lens for understanding how feedback is shaped not only by pedagogical intentions but also by institutional, relational, and emotional norms. Carless and Winstone (2020:156) define feedback cultures as 'the shared understandings, expectations and values that underpin feedback practices and interactions in a given learning context'. Rather than viewing feedback as a neutral, transactional act, this perspective highlights its social and cultural construction, rooted in power relations and institutional histories. This paper additionally situates correction-dominant feedback cultures within a wider behaviourist orientation to teaching, where feedback functions primarily as external reinforcement and error suppression, rather than as mediation for meaning-making and development.

In EFL writing education, feedback cultures are frequently characterised by correction-oriented paradigms and strong teacher authority. Ajjawi and Boud (2017: 252–260) observe that feedback is often experienced as asymmetrical, with teachers positioned as gatekeepers of correctness and learners expected to decode, accept, and comply. Such cultures privilege surface-level linguistic accuracy over rhetorical development, metacognitive awareness, and learner voice. While these dynamics are particularly visible in exam-oriented EFL contexts, they are by no means context-specific. Across many compliance-driven and assessment-intensive educational systems, feedback is institutionalised as surveillance of error rather than as dialogue about meaning-making, communicative intent, or writer identity. In these settings, visible correction practices—often symbolised by the "red pen"—come to

signify not only linguistic deviation but also academic judgement, reinforcing hierarchical feedback relations and limiting opportunities for dialogic engagement (Vlanti, 2012: 92).

Feedback cultures are also affectively charged as students interpret feedback through emotional filters shaped by prior experiences, teacher relationships, and classroom hierarchies. In high-stakes environments, such as university entrance exams or international certification tests, feedback can provoke anxiety, shame, or disengagement. Yet, despite this emotional impact, the affective dimension remains marginalised in both research and practice (Ryan & Henderson, 2018:884).

To address this gap, this paper introduces the concept of *Affective-Epistemic Feedback Habitus*: the internalised emotional and epistemological dispositions learners acquire through repeated exposure to culturally dominant feedback practices. Drawing on Bourdieu's (1977) theory of habitus, affect-informed pedagogy (Ushioda, 2011; Dewaele, 2020), and sociocultural feedback research (Carless & Boud, 2018; Ajjawi & Boud, 2017a), this concept captures how feedback is felt, trusted, and acted upon. Unlike feedback cultures, which describe external norms, or scripts, which outline expected routines, this habitus reflects a deeply embodied orientation shaping learners' intuitive reactions, compliance, resistance, anxiety, or detachment, especially in systems where correction, authority, and emotional vulnerability are intertwined.

More specifically, the *Affective-Epistemic Feedback Habitus* refers to a patterned "readiness" to interpret feedback through (a) affective expectancy (anticipating criticism, threat, or relief), and (b) epistemic positioning (treating the feedback source as unquestionable authority vs. negotiable interlocutor). In SLA terms, it helps explain why two learners receiving similar feedback may diverge sharply in uptake: one may revise strategically and reflectively, while another may either comply mechanically or disengage, depending on prior socialisation into what feedback *is* **and what it** *does* **to** the self. In EFL writing, where identity, voice, and public evaluation intersect, this habitus can amplify foreign language anxiety (Horwitz et al., 1986:125), shape motivation and self-concept (Ushioda, 2011:199), and condition whether feedback becomes "usable information" or an affective threat requiring avoidance (Ryan & Henderson, 2018: 884). The construct also clarifies a crucial pedagogical point: feedback literacy is not developed in a vacuum; it is mediated by dispositions. Where learners have internalised correction-as-judgment, they may possess procedural knowledge of "fixing errors" yet lack the epistemic confidence and emotional safety needed to negotiate feedback dialogically (Carless & Boud, 2018: 1315).

2.2 Sociocultural and Affective Pedagogy

To theorise feedback beyond static delivery, this paper draws on sociocultural learning theory, particularly Vygotsky's concepts of mediation (1978) and the Zone of Proximal Development (ZPD). From this perspective, feedback is most effective when it functions as a mediational tool, scaffolding learners' movement from current to potential performance within a supportive social context (Lantolf & Thorne, 2006:79). Unlike mechanistic correction, feedback is conceived as a dialogic process co-constructed between teacher and learner, embedded in the dynamic interplay of cognition, emotion, and social interaction.

The integration of LLM-based tools into feedback processes both aligns with and challenges this model. On the one hand, GenAI can provide timely, personalised suggestions that scaffold independent revision; on the other, its lack of relational and emotional attunement risks flattening the dialogic space, reducing feedback to surface-level edits detached from learner identity and intention. This raises critical questions about the locus of mediation, whether it resides in the algorithm, the teacher, or their interaction, and about how such mediation resonates with learners' affective and cultural expectations. It also foregrounds pragmatic concerns: culture-sensitive feedback depends not only on *what* is suggested, but

on *how* it is said, what it presupposes, and how learners interpret it as a speech act involving stance, politeness, and implied authority.

Affective dimensions—long marginalised in feedback research—are central to understanding learner engagement in EFL writing, where personal expression, linguistic vulnerability, and public evaluation intersect. Motivation and emotion shape learners' beliefs, self-concept, and agency (Ushioda, 2011: 199), while affective responses to feedback, from pride to shame, directly influence whether learners revise, persist, or disengage (Dewaele, 2020). In many EFL contexts characterised by hierarchical teacher–student relations, strong evaluative traditions and authority-driven feedback practices, teacher commentary often carries moral and identity-related weight, positioning feedback as judgment rather than guidance. In such settings, replacing human feedback with algorithmic output may weaken the relational and motivational bonds that sustain learner effort, particularly when feedback has historically functioned as a source of validation, reassurance, or control. More broadly, this concern extends to any instructional context in which feedback is treated primarily as "control information" rather than as "learning dialogue": affective safety is not a peripheral consideration but a necessary condition for sustained revision, risk-taking, and the development of authorial voice in second language writing.

To address these tensions, the paper adopts the lens of feedback literacy, understood as the capacity to understand, evaluate, and use feedback productively (Carless & Boud, 2018:1318). Feedback literacy extends beyond decoding comments to include interpreting purpose, negotiating meaning, and integrating feedback into one's learning trajectory. In AI-mediated contexts, this literacy must be explicitly cultivated: learners require guidance not only in using GenAI tools, but in critically engaging with their suggestions. Accordingly, the paper treats AI literacies as a complementary set of competencies, critical, ethical, and interactional, needed to interpret LLM outputs, recognise their limitations, and maintain authorship. Because GenAI introduces a new epistemic actor into the feedback ecology, feedback literacy must expand to include understanding probabilistic outputs and evaluating suggestions against rhetorical intent and task criteria, rather than accepting them as authoritative correction.

In AI-augmented feedback ecologies, agency must be preserved and pedagogically scaffolded. Learners should be positioned as active participants in a dialogic feedback loop, with AI as support rather than a surrogate. Teachers, in turn, act as co-mediators, helping students interpret, critique, and personalise LLM-generated feedback in ways that are both culturally and emotionally responsive. Crucially, this co-mediation is not only "pedagogical" but also "pragmatic": when LLMs produce feedback in ways that misalign with local norms of politeness, directness, stance, or implicature, learners may misread the intent or weight of suggestions. Therefore, culture-sensitive feedback design must consider the pragmatics of LLM interaction, not only its grammatical correctness.

2.3 Correction-Dominant Feedback Cultures in EFL Writing: A Global Problem with Local Manifestations

Correction-focused feedback practices remain deeply entrenched in EFL writing instruction across many educational systems, reflecting broader institutional, ideological, and assessment-driven traditions. From early schooling onward, learners in numerous exam-oriented contexts are socialised into feedback cultures where writing is evaluated primarily through annotation, error marking, and visible correction rather than dialogic engagement with meaning, voice, or rhetorical intent. These practices are not unique to any single national context; rather, they represent a widely recognisable legacy of behaviourist pedagogy in which learning is operationalised as error reduction through external reinforcement, a logic shown to constrain deep learning, undermine emotional well-being, and limit sustained writing development when it dominates feedback practices (Semke, 1984: 195; Ryan & Henderson, 2018: 884).

Within this broader landscape, EFL writing classrooms illustrate how high-stakes assessment regimes concretely shape feedback practices. In many educational systems, exam-driven curricula have historically framed writing as a test of grammatical accuracy and lexical control rather than communicative effectiveness or rhetorical development. Such regimes, characterised by standardised criteria, time pressure, and accountability structurestend to reward formulaic responses and surface-level correctness over exploratory drafting, voice, or audience awareness. Consequently, writing instruction often becomes a high-risk, performance-oriented activity in which success is defined by error minimisation rather than the cultivation of authorial agency. While the intensity of these pressures varies across contexts, their pedagogical consequences are consistently observable in EFL classrooms governed by standardisation, external evaluation, and outcome-driven assessment logics.

As a result, feedback is frequently framed as evaluative control rather than as a learning dialogue. Teachers—operating under institutional and societal pressure to deliver measurable outcomes—are positioned as grammatical arbiters whose role is to detect deviation from standard usage, mark errors (often in red pen), and return work with limited explanation or scaffolding. The red ink itself functions as a powerful pedagogical symbol, signalling judgment, finality, and correctness. Students are rarely invited into reflective dialogue about their rhetorical choices; instead, feedback is experienced as a verdict to be accepted and implemented rather than negotiated or questioned—if revision is encouraged at all (Semke, 1984: 195).

Such correction-dominant practices carry substantial affective weight. Writing becomes a high-risk act, exposing learners to correction, disappointment, and potential loss of face. Rather than cultivating revision as a recursive, exploratory process, feedback often reinforces surface-level edits and compliance with perceived norms. Over time, learners may become rhetorically risk-averse, overly dependent on memorised templates and model answers, or reluctant to develop an individual voice. The emotional cost is particularly acute for students who struggle with grammatical accuracy or who internalise early labels of being “weak” in English—labels closely associated with foreign language anxiety and diminished self-efficacy (Horwitz et al., 1986: 125).

Through repeated exposure to such feedback regimes, these tendencies gradually become dispositional rather than situational. Learners develop what this paper conceptualises as an *Affective-Epistemic Feedback Habitus*, shaped by authoritative correction, high-stakes evaluation, and emotionally charged feedback encounters. This habitus predisposes learners to interpret feedback as final, to internalise correction as personal failure, and to avoid the reflective, iterative dimensions of revision. It functions as an interpretive filter through which all feedback, whether teacher-delivered or GenAI-generated, is processed. More specifically, it normalises compliance over inquiry (epistemic stance) and threat over possibility (affective stance), making “fixing errors” feel like the only legitimate response to feedback, even when meaning-focused or rhetorical revision would be pedagogically preferable. In this sense, the red pen operates not merely as a tool but as a socialising practice: it trains learners to locate authority outside the self and to treat correctness as externally imposed truth rather than jointly constructed meaning.

The emotional climate surrounding feedback is rarely addressed explicitly in classrooms. Across many EFL contexts, teachers receive limited training in affect-sensitive pedagogy, and institutional support for formative, dialogic feedback remains constrained by curricular demands. Societal and parental expectations for academic success further reinforce accuracy-oriented norms, sustaining a pervasive climate of anxiety around writing. While such dynamics are not exclusive to EFL, they are intensified by the symbolic power of English as a gatekeeping language for educational mobility and professional opportunity (Vogt & Tsagari, 2014: 53).

The lived consequences of these feedback cultures are reflected in familiar classroom experiences. A student receives an essay densely marked with red annotations, misspellings, verb tense corrections, punctuation errors—accompanied by a grade and a single directive ("Revise"), with no comment on argument or organisation. Another learner, preparing for an international language examination, is instructed to memorise fixed essay templates and discouraged from rhetorical experimentation, receiving the implicit message that conformity, not expression, ensures success. Such vignettes, common across exam-driven EFL systems, illustrate how institutional pressure, pedagogical intent, and emotional impact converge in correction-dominant feedback cultures.

Within these contexts, GenAI-enhanced technologies introduce both risk and possibility. If learners approach GenAI primarily as a grammar checker, and if teachers implicitly legitimise this use by aligning AI feedback with error elimination rather than metacognitive engagement, GenAI risks becoming a digital red pen—efficient, immediate, and scalable, yet pedagogically reductive. This risk is not hypothetical. In accuracy-driven systems, the most socially rewarded and time-efficient use of GenAI is often surface correction ("fix my grammar," "improve my vocabulary"), because it aligns seamlessly with assessment incentives, parental expectations, and entrenched classroom routines (Vogt & Tsagari, 2014: 57).

Crucially, platform affordances can amplify this trajectory. The ease of copy–paste revision, one-click rewriting, and "polish" prompts encourages workflows where texts are edited into acceptability rather than revised into meaning, and improvement is equated with linguistic cleanup rather than conceptual rethinking. Over time, this can normalise a revision habitus of outsourcing, in which learners implement changes they cannot explain, weakening feedback literacy and diminishing the inner dialogue that makes feedback educative (Nicol, 2021:24; Carless & Boud, 2018:1318). In such ecologies, a dialogic technology is domesticated into an instrumental workflow.

Yet disruption also creates space for reconfiguration. With teacher-led mediation, explicit attention to affect and pragmatics, and the cultivation of feedback and AI literacies, LLM systems could support a shift toward process-oriented writing and dialogic revision practices that redistribute interpretive agency. Achieving this shift requires more than adopting new tools; it demands critical reflection on the historical, emotional, and institutional architectures of feedback that shape how learners interpret and respond to feedback, whether human or algorithmic. It also requires shared understandings of what constitutes meaningful AI use in revision (e.g., questioning, justifying, resisting over-editing), alongside sustained teacher professional development addressing pedagogical orchestration, ethics, and the pragmatics of human–LLM interaction.

2. LLMs and Feedback Mediation

3.1 Affordances of GenAI Feedback Tools

LLM-based tools such as ChatGPT, Claude, Gemini, and QuillBot have introduced new paradigms of writing support by delivering real-time, personalised, context-aware feedback that is typically non-judgmental in tone. Unlike traditional feedback, which is often delayed, constrained by teacher workload, and emotionally charged, these systems can address concerns ranging from grammatical accuracy to rhetorical structure and lexical sophistication (Cavaleri et al., 2024: 2; Liu & Wang, 2023: 108).

A key pedagogical affordance of LLM-mediated feedback lies in its capacity to scaffold iterative revision. Through successive prompts and paraphrasing cycles, learners can refine their writing via experimentation, reflection and revision, aligning with Vygotsky's Zone of Proximal Development, where

learning is optimally supported just beyond independent capability (Lantolf & Thorne, 2006: 80). Used intentionally, GenAI can function as a digital scaffold, enabling learners to identify gaps, explore alternatives and reflect on linguistic choices without immediate penalty.

The dialogic capabilities of advanced GenAI platforms, especially those with memory and adaptive feedback, further enable metacognitive engagement. Learners can ask why a sentence is awkward, how to strengthen an argument, or whether a word choice suits a formal register. Such "explain-this-feedback" interactions promote conceptual awareness of genre, register, and audience, key aspects of academic writing literacy, and foster what Nicol (2021: 24) terms *feedback as an inner dialogue*, where meaning is co-constructed through formative engagement.

In this light, GenAI can reposition feedback from a static, teacher-delivered product to a dynamic, learner-initiated process. When framed pedagogically, these tools have the potential to empower students as proactive agents in their writing development, using AI not as a shortcut but as a dialogic partner in composition. This dialogic potential is also pragmatic: LLMs can model stance, hedging, politeness strategies, and audience-sensitive reformulation, dimensions that matter for culture-sensitive feedback and for learners' perception of whether feedback is respectful, supportive, and actionable.

3.2 Tensions Between GenAI Feedback and Correction-Dominant Feedback Norms in EFL Contexts

Despite their dialogic affordances, the integration of LLM-mediated feedback systems into EFL writing instruction generates significant pedagogical and cultural tensions across diverse educational contexts. Central among these is a recurrent mismatch between GenAI's probabilistic, non-authoritative feedback style and entrenched correction-dominant feedback norms that characterise many exam-oriented and accuracy-driven language education systems worldwide. In such contexts, learners are often socialised to interpret feedback as a final evaluative judgment, closely associated with grades, visible correction, and error identification, rather than as an open-ended invitation to revise, reflect, or negotiate meaning.

This tension is particularly visible in EFL settings shaped by behaviourist legacies, where writing instruction has historically privileged linguistic correctness, rule adherence, and conformity to standardised norms. Within these systems, feedback is expected to provide unequivocal answers to the binary question of correctness. Consequently, GenAI's non-directive suggestions (e.g., *"You may wish to consider revising this sentence for clarity"*) may be perceived as insufficiently authoritative or pragmatically ambiguous. Empirical observations from Southern European and other exam-oriented contexts indicate that learners frequently seek categorical confirmation when encountering AI feedback, asking variations of "Is this right or wrong?" (Garcia & Lee, 2023: 89). This response signals a deeper epistemological tension: GenAI's context-sensitive, probabilistic recommendations contrast sharply with the rule-based certainties cultivated by traditional EFL feedback regimes.

In many EFL contexts internationally, these tensions are particularly salient. Learners are often socialised within highly evaluative feedback cultures in which teacher authority, visible correction practices (e.g., red-pen marking), and high-stakes assessment structures dominate writing instruction. For learners shaped by an *Affective-Epistemic Feedback Habitus* that equates feedback with authoritative correctness, GenAI's mitigated, probabilistic guidance may be experienced as confusing, insufficiently decisive, or even unreliable unless explicitly mediated by teachers. Importantly, these dynamics are not confined to a single national setting. Comparable patterns have been documented across diverse EFL contexts where assessment incentives, institutional accountability, and parental expectations prioritise surface accuracy

and error elimination over rhetorical development, authorial agency, and dialogic engagement with feedback.

Recent scholarship on the pragmatics of LLM-mediated interaction further complicates assumptions that AI-generated feedback is neutral or universally interpretable. Research in computational pragmatics and AI discourse analysis demonstrates that LLMs systematically encode politeness strategies, epistemic hedging, indirectness, and face-preserving formulations rooted primarily in Anglophone academic discourse norms (Brown & Levinson, 1987: 66–68; Danescu-Niculescu-Mizil et al., 2013: 8; Hovy & Spruit, 2016: 592). While such pragmatic features may soften affective impact, they also introduce interactional ambiguity in educational cultures accustomed to directive, authority-driven feedback. LLM-generated feedback typically privileges mitigated suggestions over categorical correction, reflecting probabilistic alignment rather than pedagogical intentionality (Bender et al., 2021: 615; Floridi et al., 2018: 694). In many EFL contexts, this pragmatic mismatch can lead learners either to dismiss AI feedback as indecisive or to overinterpret it as covertly authoritative, thereby reinforcing compliance rather than dialogic engagement. These dynamics underscore the need to conceptualise GenAI feedback not only as a cognitive or affective intervention, but as a socio-pragmatic act whose interpretation is culturally mediated and pedagogically consequential.

These tensions are further intensified by what may be described as feedback shortcutting. Learners may rely on tools such as Grammarly or ChatGPT to eliminate surface errors while bypassing the cognitive work of developing writing strategies. In assessment-driven EFL systems, where correctness is rewarded more visibly than rhetorical or metacognitive growth, this reliance risks reducing feedback to linguistic outsourcing (Vogt & Tsagari, 2014:57). This mechanism explains how an intuitively dialogic technology can become a digital red pen: dominant success metrics (error-free text) steer usage toward the fastest visible gains (surface correction), reinforcing compliance-based revision habits and weakening reflective uptake. Moreover, a subtle standardisation effect emerges, as LLMs, optimised for fluent, norm-aligned prose, implicitly privilege conventional phrasing and generic correctness, especially when learners request *"better"* or *"more advanced"* English without articulating rhetorical intent. Over time, this norm-alignment risks reproducing correctional hierarchies under a new technological interface, positioning learners as implementers of external edits rather than negotiators of meaning (Ajjawi & Boud, 2017b: 252–260; Carless & Winstone, 2020:15 6).

This risk is amplified by techno-solutionist discourses that frame GenAI as an efficiency-enhancing remedy for long-standing feedback challenges such as teacher workload, delayed correction, or error density. When adopted primarily as a tool for speed, scalability, or cost reduction, GenAI may automate and intensify existing correctional paradigms rather than transform them. In EFL systems shaped by high-stakes assessment, techno-solutionist adoption often positions GenAI as a mechanism for faster error detection and surface-level optimisation, thereby reinforcing behaviourist feedback logics under the guise of innovation. Without deliberate pedagogical reframing, GenAI does not disrupt the red-pen culture; it digitises it.

Teacher roles are also implicated in this shift. If LLM-powered feedback is positioned as a substitute for teacher input, educators risk marginalisation, with potential erosion of both pedagogical authority and the relational trust that sustains learner motivation (Semke, 1984:197). In the absence of structured mediation, LLM-generated feedback may operate in a pedagogical vacuum, technically proficient yet affectively and culturally misaligned. This misalignment is not only emotional but pragmatic: LLM feedback may mishandle politeness, directness, implicature, humour, or culturally preferred degrees of explicitness, leading learners to misinterpret suggestions as overly authoritative, insufficiently clear, or excessively interventionist. Culture-sensitive integration, therefore, requires sustained attention to the

pragmatics of human–LLM interaction, including how stance, interpersonal alignment, and evaluative force are communicated through AI feedback.

Addressing these tensions necessitates embedding LLM-mediated feedback within teacher-facilitated, dialogic feedback cultures. Learners require explicit support in developing critical feedback literacy: the capacity to interpret, critique, adapt, and, when appropriate, reject algorithmic suggestions in line with rhetorical purpose, audience awareness, and communicative intent. Such discernment ensures that GenAI functions as a dialogic support rather than a decontextualised corrector, preserving both pedagogical integrity and learner agency. Crucially, this training must be framed as part of AI literacies alongside feedback literacy, encompassing interpretive competencies (evaluating output quality), interactional competencies (formulating productive prompts), and ethical competencies (maintaining authorship and academic integrity). In practice, this involves designing feedback tasks that make *thinking with feedback* visible, requiring justification, annotation, and comparison, so that the learning outcome is not merely a cleaner text, but a more agentive, reflective writer. Taken together, these dynamics illustrate how GenAI's pedagogical impact is not determined by its dialogic capacity alone, but by the feedback cultures into which it is embedded.

3.3 Emotional and Motivational Implications

One of the most under-theorised yet critical dimensions of GenAI feedback is its emotional impact on learners. In feedback cultures where fear of judgment, performance anxiety, and correction fatigue are prevalent, LLM-mediated feedback offers a potential affective shift. By presenting suggestions in a neutral, impersonal tone, GenAI may reduce the emotional sting often associated with teacher correction. For students who link teacher feedback with failure or embarrassment, AI can provide a psychologically safer space for experimentation and risk-taking in writing (Ryan & Henderson, 2018: 884). Emotional safety in this context is not trivial: it underpins creativity, cognitive engagement, and willingness to take linguistic risks, particularly in writing, where personal expression meets linguistic vulnerability.

When framed pedagogically, GenAI can help learners move beyond formulaic structures toward more autonomous and confident expression. This aligns with Ushioda's (2011:202) view of motivation as a relational and affective construct shaped by the learner's emotional experience of instructional activities and evaluative interactions.

Yet the absence of affectively salient human feedback carries risks. *Authorship blurring*—the uncritical incorporation of AI suggestions—may dilute ownership of the text, complicating questions of originality and reducing opportunities for metacognitive reflection. Learners may revise "correctly" without understanding why, resulting in superficial gains. Overreliance on GenAI can also foster motivational detachment: if revision is routinely outsourced, writing may be perceived less as personal expression or growth and more as a mechanical process to be optimised algorithmically. This instrumentalisation of writing risks weakening intrinsic motivation and eroding confidence in one's authorial voice.

To mitigate these risks, GenAI must be positioned not as an evaluator but as a conversational partner whose feedback is to be interpreted, adapted, or sometimes rejected. Teachers should help learners develop discernment, ownership, and critical trust in both human and AI feedback. Without such mediation, the promise of transforming feedback into a dialogic, learner-centred process may remain unrealised, particularly in cultures where feedback is historically bound to emotional vulnerability and academic authority. From a pragmatic perspective, this also requires attention to how LLM feedback performs stance: hedges ("perhaps," "consider"), directives ("change X to Y"), and evaluative language can be interpreted differently across cultures and proficiency levels, affecting whether learners perceive feedback as supportive, ambiguous, or authoritative.

4. Reimagining Feedback Cultures: Pedagogical Implications

In many EFL contexts, these tensions become particularly visible. Learners are often socialised within highly evaluative feedback cultures dominated by teacher authority, red-pen correction, and high-stakes assessment. For learners shaped by an *Affective-Epistemic Feedback Habitus* that equates feedback with authoritative correctness, GenAI's mitigated or non-directive guidance may prompt confusion, mistrust, or dismissal unless explicitly mediated by teachers. Comparable dynamics have been documented across EFL settings where institutional expectations, assessment incentives, and parental pressures prioritise surface accuracy over rhetorical development and dialogic engagement.

Rather than treating LLMs as autonomous evaluators, educators should embed them within frameworks that protect learner agency, emotional safety, and cultural specificity. Unmediated use risks automating existing correctional paradigms; thoughtful integration can instead position LLMs as dynamic mediators of writing development, prompting reflection, negotiation, and ownership. Achieving this transformation hinges on two interdependent priorities: repositioning teachers as feedback mediators and cultivating learner feedback literacy in LLM-rich environments. A third, cross-cutting priority is the explicit development of AI literacies and sustained teacher professional development, ensuring that AI does not function as an invisible curriculum silently reshaping what counts as writing, revision, and improvement.

4.1 Pedagogical Repositioning of Teachers

The introduction of GenAI into feedback processes necessitates a shift in the teacher's role—from evaluator to feedback mediator. Teachers are called to move beyond acting as final arbiters of correctness and instead function as interpreters and co-constructors of meaning, guiding learners through GenAI outputs with critical, affective, and contextual awareness (Winstone & Carless, 2020: 67).

This repositioning does not diminish teacher expertise; rather, it amplifies it by foregrounding the orchestration of meaningful interaction between learner and machine. Effective mediation involves helping students understand not only what GenAI suggests, but why, how, and when those suggestions align, or conflict, with communicative goals, genre conventions, and audience expectations, particularly when AI feedback is vague, formulaic, or culturally misaligned. In this way, teachers preserve feedback as dialogue, ensuring that AI functions as a catalyst for reflection rather than a decontextualised corrector (Nicol, 2021:27). Crucially, such mediation also supports pragmatic alignment, enabling learners to interpret stance, hedging, directness, and politeness in AI feedback and to negotiate revisions that remain rhetorically and culturally appropriate.

Teachers must also attend to the emotional dimension of AI-mediated feedback, supporting learners who may experience confusion, dependency, or diminished authorship when working with algorithmic suggestions. This affective scaffolding is essential for sustaining motivation and trust—elements often absent from purely algorithmic exchanges (Ryan & Henderson, 2018:885). Teacher professional development is therefore central: educators require not only technical familiarity with GenAI tools, but also AI literacies, critical pedagogical frameworks, and guidance in orchestrating dialogic human–AI–learner feedback interactions.

4.2 Cultivating Feedback Literacy in AI-Mediated Settings

To unlock the transformative potential of LLMs in writing instruction, feedback must be reimagined not merely as correction, but as a literacy—an evolving set of interpretive, reflective, and dialogic

competencies that learners actively develop. Feedback literacy, as defined by Carless and Boud (2018: 1319), involves understanding feedback processes, recognising its role in learning, and acting on it effectively. In GenAI-mediated classrooms, this literacy must extend to *algorithmic feedback discernment*.

Accordingly, feedback literacy must be complemented by *AI literacy:* the capacity to understand how GenAI systems generate feedback, recognise their probabilistic nature and limitations, and distinguish algorithmic suggestions from pedagogically intentional guidance. AI literacy intersects with, but does not replace, feedback literacy. Without this layered competence, learners risk treating pragmatically hedged, probabilistic AI output as authoritative correction, defaulting to passive uptake rather than reflective engagement.

Developing authorship awareness, therefore, requires learners to interpret, negotiate, and selectively adopt GenAI feedback. Although LLMs can provide flexible, context-sensitive suggestions, their outputs may be generic, misaligned with rhetorical intent, or culturally inappropriate. Learners must also develop pragmatic discernment: recognising when AI reformulations alter stance, modulate politeness, intensify or soften claims, or disrupt discourse expectations relative to local classroom norms, exam genres, or target communities of practice.

Crucially, this process is mediated by the *Affective-Epistemic Feedback Habitus* learners bring to the classroom. Where learners have been socialised into authority-driven, correction-oriented feedback regimes, dispositions of deference, anxiety, and compliance may lead to uncritical acceptance of AI suggestions, undermining dialogic engagement unless explicitly addressed.

To embed LLMs into feedback processes rather than products, teachers can design tasks that require learners to compare AI suggestions with original intentions, justify acceptance or rejection of revisions, annotate changes using genre- or register-based rationales, and reflect on how feedback reshapes rhetorical stance. Such practices position GenAI feedback as a catalyst for reflection rather than an endpoint for correction, aligning with constructivist principles of learner agency.

Finally, explicit classroom discussions about AI's limits, its biases, lack of socio-pragmatic awareness, and tendency toward over-editing can demystify the tool and foster critical digital literacy. Students thus come to see AI not as an infallible authority, but as a helpful, fallible, and negotiable feedback partner. These discussions should include "LLM pragmatics": how prompts function as speech acts, how politeness and stance are conveyed in AI feedback, and how culturally variable interpretations of hedging, directness, and evaluation can shape learner uptake. This is where AI literacies intersect with feedback literacies: the goal is not merely competent use of tools, but informed judgment about when AI supports learning and when it substitutes for it.

4.3 Localising GenAI Use in Writing Feedback Across Educational Contexts

Crucially, the successful integration of LLMs into feedback practices requires culturally responsive adaptation across educational contexts. Writing classrooms—across EFL, ESL, and academic literacy settings—operate within sociocultural ecosystems where feedback is tightly intertwined with identity, authority, evaluation, and emotional exposure. Ignoring these dynamics in favour of universalised "best practices" risks reproducing the very disconnects that have historically limited the effectiveness of feedback in diverse educational systems (Vogt & Tsagari, 2014: 59). Localisation, therefore, should not be misconstrued as parochialism: it constitutes a core design principle for responsible GenAI integration, precisely because feedback is always situated, even when the technologies that deliver it are global.

Accordingly, GenAI must be localised rather than merely adopted. This entails mediating LLM use in ways that account for learners' linguistic profiles, rhetorical traditions, and socio-affective expectations. Across many educational contexts, learners bring specific patterns of L1 transfer, genre socialisation, and assessment-oriented writing habits that shape how feedback is interpreted and acted upon. AI-generated feedback that simply flags errors or rewrites text without contextual explanation risks reinforcing surface-level correction, regardless of the language or setting. Educators, attuned to both the linguistic and emotional realities of their classrooms, are uniquely positioned to provide this mediation, ensuring that AI feedback supports learning rather than compliance. Localisation must also involve pragmatic calibration: expectations surrounding directiveness, mitigation, praise, and critique vary widely across educational cultures, and LLM outputs may inadvertently signal excessive authority, unwarranted tentativeness, or ambiguity. Attending the pragmatics of LLM feedback is therefore integral to culture-sensitive feedback design.

At the same time, GenAI integration must actively resist techno-solutionism, the assumption that automation, speed, or efficiency alone can resolve pedagogical challenges. When LLMs are introduced primarily to address teacher workload, turnaround time, or error density, they **risk** digitising correctional paradigms rather than transforming them. In exam-oriented and accountability-driven systems worldwide, such adoption frames GenAI as a tool for rapid surface-level optimisation, thereby reinforcing behaviourist feedback logics under the guise of innovation. Teacher-guided adaptation is therefore essential: educators must deliberately curate how and when GenAI is used, monitor its effects on learner motivation and revision behaviour, and embed it within feedback cycles that privilege process, interpretation, and reflection over performance outcomes.

Addressing these risks requires sustained teacher professional development that extends well beyond technical training. Educators need structured support in critical AI literacy, pedagogical mediation, and affect-sensitive feedback design, enabling them to interpret GenAI outputs, anticipate learner misinterpretations, and integrate AI feedback into dialogic instructional routines rather than allowing it to function as a parallel or substitute authority. Professional development should explicitly address AI literacies, including tool evaluation, prompt pedagogy, authorship, and ethical considerations, as well as pragmatic awareness, such as how AI feedback communicates stance, politeness, and implied authority across contexts.

Reimagining feedback cultures through GenAI, therefore, demands more than functional integration; it requires a cultural, emotional, and pedagogical recalibration across educational systems. Teachers must be repositioned as mediators of meaning, learners cultivated as literate feedback agents, and AI grounded in the socio-affective realities of local classrooms. Only under these conditions can feedback evolve from correction to collaboration, from surveillance to scaffolding, and from red ink to intelligent dialogue.

5. Conclusion and Future Directions

As LLM-mediated tools increasingly permeate educational settings, their implications for feedback practices in EFL writing demand urgent theoretical and pedagogical scrutiny. This article has argued that GenAI is not a neutral instructional aid, but a socio-technical actor whose effects are shaped by the pedagogical, cultural, and institutional ecologies into which it is introduced. Across EFL contexts, GenAI can either reinforce entrenched correction-dominant paradigms or contribute to the reconfiguration of feedback cultures as dialogic, reflective, and learner-centred processes. Crucially, where feedback has historically been organised around behaviourist logics of error detection, compliance, and external evaluation, uncritical adoption of GenAI risks assimilating new technologies into existing correctional regimes rather than transforming them.

A central risk identified in this paper is techno-solutionism: the assumption that the introduction of advanced technologies will, in and of itself, resolve long-standing pedagogical problems related to feedback quality, learner engagement, or teacher workload. When GenAI is adopted primarily for its efficiency, speed, or scalability, without explicit pedagogical reframing, it risks functioning as a *digital red pen*, automating correction-as-judgment and intensifying compliance-oriented revision practices through algorithmic means. Techno-solutionism thus represents not a separate issue but a *mechanism of reinforcement*: it enables behaviourist feedback logics to persist under the appearance of innovation, particularly in assessment-driven EFL systems where surface accuracy remains the most visible indicator of improvement.

The analysis underscores that resisting techno-solutionism requires more than caution; it requires explicit cultivation of literacies. First, feedback literacy is essential if learners are to engage with GenAI feedback as an interpretive resource rather than a verdict. Without the capacity to question, negotiate, and justify revisions, learners may default to passive uptake, reproducing epistemic dependency rather than developing authorial agency. Second, AI literacies are indispensable to understanding how LLMs generate feedback, recognising probabilistic and pragmatically hedged outputs, and evaluating suggestions against rhetorical intent, task criteria, and audience expectations. These literacies are not peripheral but foundational: without them, GenAI risks becoming an invisible curriculum that silently reshapes what counts as "good writing" and "successful revision."

Equally critical is the role of teacher professional development in mediating these processes. Teachers remain central to preventing uncritical GenAI adoption, not only by guiding tool use but by framing feedback as a dialogic, affectively safe, and culturally situated practice. Professional development must therefore extend beyond technical training to include critical AI literacy, pedagogical mediation, ethical awareness, and pragmatic sensitivity to how AI feedback communicates stance, authority, and evaluation. Without such preparation, educators may inadvertently legitimise correction-only uses of GenAI or cede epistemic authority to algorithmic outputs, allowing techno-solutionism to reconfigure, not reduce, existing feedback asymmetries.

This theoretical inquiry opens multiple pathways for future empirical research across EFL contexts. There is a pressing need for longitudinal studies examining how AI-mediated feedback shapes revision depth, authorship awareness, emotional engagement, and metacognitive development over time. Further research should also investigate pragmatic alignment in LLM feedback, specifically how learners interpret hedging, directiveness, politeness, and stance, and how these cues interact with local norms of authority and evaluation. Importantly, future studies should foreground student voice, exploring how learners experience GenAI feedback emotionally and epistemically, and how these experiences influence uptake, resistance, or overreliance.

The integration of GenAI into EFL writing education should therefore be approached not as a technological upgrade but as a pedagogical and ethical inflection point. Whether GenAI amplifies dependency or fosters autonomy, entrenches correctional authority or redistributes interpretive agency, depends less on algorithmic sophistication than on the literacies, mediation practices, and professional judgments that surround its use. Feedback ecosystems must be intentionally designed, where teachers mediate, learners interpret, and AI functions as dialogic support rather than digital overseer.

Only under these conditions can red ink be transformed into resonance and correction into collaboration. Ultimately, reshaping feedback in EFL education requires not only new tools but the unlearning of the established *Affective-Epistemic Feedback Habitus*, a pedagogical challenge that demands cultural sensitivity, teacher mediation, and dialogic reorientation. The challenge ahead is therefore not merely to

teach with AI, but to teach against the grain of inherited feedback cultures, crafting classroom ecosystems where both human and algorithmic feedback can be sources of empowerment, not surveillance. The question we must now confront is simple, yet consequential: Will we use GenAI to reinforce the red pen or to rewrite the culture of writing itself?

Limitations

This article is positioned as a theoretical and critical discussion rather than an empirical investigation or systematic review. As such, its claims are conceptual and interpretive, grounded in established scholarship, theoretical synthesis, and contextual analysis rather than primary data collection. The illustrative examples employed serve explanatory purposes, not evidentiary ones. Accordingly, the arguments advanced here should be understood as a theoretically informed agenda for future research and pedagogical reflection, intended to be empirically examined, refined, and contextualised through methodologically rigorous studies in EFL writing classrooms and comparable educational settings.

References

Ajjawi, R., & Boud, D. (2017a). Exploring the tensions of giving and receiving feedback in higher education. *Studies in Higher Education, 42*(4), 584–597. https://doi.org/10.1080/02602938.2015.1102863

Ajjawi, R., & Boud, D. (2017b). Researching feedback dialogue: An interactional analysis approach. *Assessment & Evaluation in Higher Education, 42*(2), 584–597.252–265. https://doi.org/10.1080/02602938.2015.1102863

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Bourdieu, P. (1977). *Outline of a theory of practice*. Cambridge University Press.

Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage*. Cambridge University Press.Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enablinguptake of feedback. *Assessment & Evaluation in Higher Education, 43*(8), 1315-1325. https://doi.org/10.1080/02602938.2018.1463354

Carless, D., & Winstone, N. (2020). Teacher feedback, literacy, and student engagement with feedback: Reviewing the evidence. *Educational Research Review, 31*, 100326. https://doi.org/10.1016/j.edurev.2020.100326

Cavaleri, M., Jenkins, K., & Somasundaram, J. (2024). Harnessing generative AI for academic writing: Affordances, anxieties, and pedagogical strategies. *Computers and Composition, 69*, 102801. https://doi.org/ 10.58723/jaiela.v1i1.56

Danescu-Niculescu-Mizil, C., Sudhof, M., Jurafsky, D., Leskovec, J., & Potts, C. (2013). A computational approach to politeness with application to social factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 250–259). Association for Computational Linguistics.

Dewaele, J.-M. & Li, C. (2020). Emotions in Second Language Acquisition: A Critical Review and Research Agenda. *Foreign Language World*, (1): 34-49.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines, 28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Garcia, A., & Lee, M. (2023). Learning with machines: Student attitudes and practices in AI-supported writing. *Language Learning & Technology, 27*(2), 85–97.

Horwitz, E. K., Horwitz, M. B., & Cope, J. (1986). Foreign language classroom anxiety. *The Modern Language Journal, 70*(2), 125–132.

Hovy, D., & Spruit, S. L. (2016). The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 591–598). Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-2096

Lantolf, J. P., & Thorne, S. L. (2006). *Sociocultural theory and the genesis of second language development*. Oxford University Press.

Liu, X., & Wang, H. (2023). Generative AI and writing development: From language correction to authorship support. Language Learning & Technology, 27(1), 103–121.

Lyrigkou, C. (2021). *The role of informal second language learning in the spoken productions of EFL learners in Greece* (Doctoral dissertation). The Open University, UK.

Nicol, D. (2021). The power of internal feedback: Exploiting natural comparisons in assessment design. *Assessment & Evaluation in Higher Education, 46*(3), 450–462. https://www.tandfonline.com/doi/full/10.1080/02602938.2020.1823314

Ryan, T., & Henderson, M. (2018). Feeling feedback: Students' emotional responses to educator feedback. *Assessment & Evaluation in Higher Education, 43*(6), 880–892. https://www.tandfonline.com/doi/full/10.1080/02602938.2017.1416456

Semke, H. D. (1984). Effects of the red pen. *Foreign Language Annals, 17*(3), 195–198.

Ushioda, E. (2011). Language learning motivation, self, and identity: Current theoretical perspectives. *Computer Assisted Language Learning, 24*(3), 199–210. https://doi.org/10.1080/09588221.2010.538701

Vlanti, S. (2012). Assessment practices in the English language classroom of Greek Junior High School. *Research Papers in Language Teaching and Learning, 3(*1), 92–122

Vogt, K., & Tsagari, D. (2014). Assessment literacy of foreign language teachers: Findings of a European study. *Language Assessment Quarterly, 11*(4), 374–402. https://doi.org/10.1080/15434303.2014.960046

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Winstone, N. E., & Carless, D. (2020). *Designing effective feedback processes in higher education: A learning-focused approach*. Routledge.

Ioanna Nifli-Sakali (iniflis@enl.auth.gr) is a Greek-Canadian English language educator, translator, and Ph.D. candidate at the School of English, Aristotle University of Thessaloniki. She holds an MA in TESOL and an MA in Applied Translation Studies from the University of Leeds. Her professional experience includes high-stakes institutional work with international organisations such as the United Nations and the European Parliament, supporting complex policy-driven communication across multilingual environments. Her doctoral research in Computational Linguistics examines the pedagogical implications of Large Language Models (LLMs) and Automated Essay Scoring (AES) systems in Greek EFL writing contexts, with broader interests in AI-mediated language learning, multilingual communication, and responsible educational innovation.

Research Papers in Language Teaching and Learning
Vol. 16, No. 1, March 2026, 47-59
ISSN: 1792-1244
Available online at http://rpltl.eap.gr

Why AI Literacy Matters in EAP: Lessons from Engineering Students' Homework Practices

Sonia Carmen Munteanu

The rapid proliferation of generative artificial intelligence (GenAI) in higher education has raised urgent questions about AI literacy and its integration into disciplinary curricula. This study examines how engineering students at a Romanian university engaged with GenAI tools when completing English for Academic Purposes (EAP) homework, in a context without institutional guidance, formal training, or clear curricular provisions. Framed within an action research approach, the study draws on student self-reports and survey responses to explore both practices and perceptions. The findings indicate that most students relied on AI tools—particularly ChatGPT—to refine language, clarify ideas, and enhance coherence in their writing. Students generally viewed these tools positively, valuing their contribution to efficiency, comprehension, and preparedness for future professional contexts. However, only limited evidence emerged of more advanced practices such as critical evaluation or iterative prompting. A smaller group preferred to work without AI, motivated by ethical concerns, self-reliance, or scepticism. The study argues for embedding AI literacy into EAP courses to foster equitable access, critical and ethical use, and pedagogical alignment with the evolving technological landscape.

Keywords: English for Academic Purposes (EAP), Generative artificial intelligence, AI literacy, Higher education, Action research

---

1. Introduction

The recent widespread adoption of GenAI tools for language learning in HE has spurred research focusing on its impact on curriculum design, teaching, learning and assessment (Lee et al., 2024). In the fields of professional communication and English for Academic Purposes (EAP), research shows that GenAI can enhance learning processes (Nelson et al., 2025) and autonomous learning (Ingley & Pack, 2023), can assist with personalized feedback (Mahapatra, 2024; Su et al., 2023), and can scaffold the development

of professional communication skills (Godwin-Jones, 2022; Khodi & Curle, 2025; Mahapatra, 2024; Su *et al*., 2023).

Since the effective use of GenAI requires a deep understanding of its complex capabilities and prompting skills to leverage its full potential, HE institutions (HEIs) have to face a multifaceted challenge when adopting it for learning and teaching. Apart from the need to provide access to it for educators and students, HEIs need to provide regulations, training and awareness of capabilities, of potential and of contradictions (Chiu et al., 2023; Darvin, 2025; Warschauer et al., 2023; Yang et al., 2025). The urge to integrate AI across the curriculum (Southworth et al., 2023) spurs changes at disciplinary level, leading to research into innovative practices that integrate AI literacy with disciplinary skills. Nonetheless, both at programmatic and implementation levels, HEIs display a varied pace in change, due to lack of institutional policies or slow curriculum updates (Vettori & Warm, 2025). More and more, however, have started to launch initiatives and programs for AI literacy development that aim to equip students and staff with foundational knowledge about AI technologies, ethical considerations, and effective, responsible use of Gen AI tools (Vettori & Warm, 2025).

Against this general background, the present study examines how engineering students in a Romanian HEI engage with GenAI tools for EAP learning within a context of vague curricular provisions, no institutional guidelines, and no formal training on their pedagogical use. The study aims to assess local conditions that could justify the inclusion of AI literacy components in the EAP curriculum by investigating two key aspects: students' reported uses of GenAI when completing homework and their perceptions of its benefits and drawbacks for language learning.

2. AI Literacy as Component of Disciplinary Curriculum

The rapid expansion of generative artificial intelligence (GenAI) in educational contexts has intensified calls for the systematic development of AI literacy as a core competency for learners. Warschauer *et al*. (2023) offer a comprehensive conceptualisation, identifying four interrelated competency areas: understanding AI (grasping core concepts, capabilities, and limitations), using AI (applying tools effectively in varied contexts), evaluating AI (critically assessing outputs for accuracy, bias, and appropriateness), and creating AI (designing or adapting systems or outputs for specific needs). This framework positions AI literacy as an interdisciplinary capability that integrates technical knowledge with cognitive, socioemotional, and ethical competencies, and it stresses the importance of equitable access to avoid exacerbating the digital divide. Education systems, the authors argue, should embed AI literacy through scaffolded, domain-specific instruction that develops learners' abilities progressively from basic familiarity to critical, reflective, and creative engagement, while recognising it as a lifelong learning process.

Complementing this broader conceptual framing, Chan and Hu's study (2023) captures higher education students' perspectives on using generative AI for learning and assessment in HE. Participants generally valued GenAI for its capacity to generate ideas, refine writing, and improve efficiency, particularly in language-related tasks. However, the study found considerable variation in how critically and ethically students engaged with AI outputs: while some used them as a springboard for further refinement, others relied on them with minimal modification. These findings underscore the need for explicit guidance and structured support in developing the critical evaluation and ethical use competencies central to AI literacy.

In the context of EAP and professional communication, the ever evolving affordances of Gen AI to produce text in multiple languages, to translate, to comprehend and control content, to adapt messages to a variety of audiences, and respond to numerous communicative demands in diverse contexts raised

concerns about language learners' use of such tools as substitutes for learning rather than assistance to learning.

Recognising the overlap between these skills typically taught in EAP courses and the capabilities of AI tools, Ngo and Hastie (2025) propose a framework for teaching AI literacy within EAP contexts. Building on this evidence of variable engagement and the need for structured support, Ngo and Hastie offer a more targeted definition of AI literacy relevant to EAP and English for Specific Purposes (ESP): AI literacy is "the ability to critically evaluate, use, and create AI applications effectively and ethically, supported by the necessary knowledge, skills, and awareness." This multidimensional framing emphasises three key competencies: critically evaluating AI output for reliability, bias, and appropriateness; using AI effectively and ethically in context; and creating or adapting AI applications to meet specific needs. While the last dimension is less immediately relevant to most EAP settings, the first two directly align with the communicative and academic demands of higher education. From an EAP and ESP perspective, AI literacy enables learners to engage with AI-generated text in ways that strengthen both language proficiency and disciplinary communication skills, provided they are explicitly taught to interrogate AI outputs, apply them appropriately, and acknowledge their use ethically.

The importance of this critical and ethical engagement is underscored by the findings of Nelson et al. (2025), which examined how learners use AI tools in academic writing tasks. The study found that while many students appreciate GenAI's ability to generate ideas, improve structure, and enhance language accuracy, their engagement often remains passive, with outputs accepted at face value and minimal evidence of systematic evaluation or adaptation. Ethical dimensions, such as correct attribution and awareness of AI's limitations, were inconsistently addressed, reflecting a lack of structured pedagogical support for these competencies.

Taken together, these studies highlight a significant gap between students' current AI practices and the fuller scope of AI literacy as conceptualised by Ngo and Hastie. Both point to the risk that, without deliberate and scaffolded instruction, learners will remain at a surface level of AI use, missing opportunities for deeper, more critical engagement. For EAP, this gap has direct implications: embedding AI literacy into the curriculum can ensure that students develop not only the technical capacity to use AI but also the reflective, ethical, and adaptive skills necessary for effective academic and professional communication in AI-rich environments.

Investigating a decade of AI literacy education, Yang et al (2025) underscored the need to embed AI literacy into curricular standards. Their study reveals that, for educational practices, one of the most effective ways of supplying AI literacy is creating opportunities for learners "to engage in hands-on activities and real-world application that enhance [their] ability to apply AI content in their lives and future career" (p.9). Thus, learners develop agency and can become responsible and informed AI users. In light of these findings and frameworks, the present study seeks to explore how such principles might be meaningfully applied in a local EAP context. By investigating engineering students' reported uses of GenAI for academic writing and their perceptions of its benefits and limitations, this research aims to inform the potential integration of AI literacy into the curriculum as a structured, pedagogically grounded practice.

3. Methodological Approach

3.1 Action Research Perspective

The main objective of the present study is to better understand how students use GenAI tools for independent work in their language learning, by analysing data collected from students I taught, within the context of an EAP curriculum I designed and implemented. Consequently, my work is situated in the tradition of action research (AR) in language education, a process of action and reflection on one's own teaching practices for the purpose of identifying what affects these practices and how they can be improved (Banegas & Consoli, 2020). Burns (2016) highlights the self-reflective, critical and systematic nature of AR, which renders it particularly suitable for practitioners who need to explore the impact of their pedagogical decisions on student learning and on educational processes so that they can take corrective measures or validate a course of action in their teaching.

Action research relies on three key elements: context, agents and issues. In language education, the context of action research may include, among others, the institution, the curriculum, various regulations and conditions that influence teachers and learners. These have "strong potential to shape and dynamically influence the practices of teaching and learning" (Banegas & Consoli, 2020: 177). The agents are understood as the active participants, who include the teachers and the learners. Teacher agency is apparent not only in their teaching and research but also in their engagement in reflection and self-reflection at every step of the research process. Action research in second/foreign language education concerns issues pertaining to learners' development of communicative competence or to teachers' professional development (Dikilitas & Griffiths, 2017).

The present study is set against the context of a HE institution that offers EAP courses to second-year engineering students, as part of their mandatory curriculum. The course spans over fourteen weeks with a two-hour weekly face-to-face session. Class sizes are large and vary between eighty to one hundred students per class. These conditions pose serious constraints on two important aspects of learning: student engagement in class and individual feedback. The course requires students to do two graded assignments as homework. Students are encouraged to work autonomously. Direct, formative feedback before submission of homework is impossible due to the same contextual constraints mentioned above. Course requirements do not include specific provisions for the use of GenAI tools. General provisions on plagiarism and other unethical behaviour apply, however. Writing assistance tools such as grammar, style or punctuation checkers or automatic correction have never been banned or restricted for homework completion.

The use of AI tools with enhanced performance in language comprehension and text production remains somewhat unacknowledged in the particular institutional context. Since the advent of LLM-based chatbots such as ChatGPT-3, their presence as actors involved in educational processes (planning curricula, teaching, learning and assessment activities) has become more frequent at all levels. Institutions, however, are slower at embedding the multitude of uses into formal policies and procedures, often limiting their efforts to preventing the potential negative impact of GenAI, rather than providing guidelines, instruction or training to teachers and learners on how to use these fast-developing technologies to enhance learning and teaching in ethical and effective ways.

Employing an AR-grounded approach, the study collected quantitative and qualitative data from the engineering students who took the EAP course in order to address the following research questions (RQ):

- RQ1: How do students report using GenAI tools for completing EAP homework assignments?
- RQ2: What are the students' perceptions of the benefits and drawbacks of using GenAI tools for EAP homework?

3.2 Participants and Data

The participants in this study are second-year engineering students who took the EAP course in the summer semester of the 2024/25 academic year. A total of 160 students (NT=160) did their homework, filled in the AI Tools Declaration of Use (AID) and submitted it with the assignment. This was mandatory, and the declarations were signed by students. From the total number, 23 declared they had not used any AI tool while doing the assignment (NT0=23), while 137 mentioned the use of AI tools for completing the assignment (NT1=137). After the final submission, all students were invited to complete a survey, which was anonymous. Only 49 students responded to this post-activity survey (SR=49), which represents 30.69% of the total number who did their homework.

The data collected consist of two types of student responses: the AIDs (NT = 160) and the survey responses (SR= 49). Both sources are multiple-choice questionnaires which can provide insights from a quantitative perspective. The post-activity survey also contained an open-ended question, which provides qualitative data. The quantitative data are analysed here using descriptive statistics, while the qualitative data have undergone a thematic analysis. Both sources contribute synergistically to a better understanding of the students' perspective on using AI tools for learning and doing homework in EAP.

The first data collection tool was a three-section document signed by the students and submitted together with the homework assignment (see Annex). The first question asked students to state whether they used any AI tool when doing the homework. Those whose answer was yes were asked to move on to the next section. This part focused on the tool used, the purpose and method of use, how the students' work was integrated with the AI contribution and how the students validated the final version before submission. The final part of the form required the students to acknowledge the ethical implications of working with AI and confirm that the work was original and a personal contribution.

The second data collection tool was a survey with seven statements to be evaluated on a five-point Likert scale (strongly agree to strongly disagree) and one open-ended question. The purpose of the questionnaire was to investigate students' perception of using AI tools for EAP homework and was adopted from the AI Motivation Scale (AIMS) instrument by Li et al. (2025). Their AIMS questionnaire is a multi-dimensional instrument designed and validated for gauging students' motivation to learn with AI in higher education. My adaptation consisted of selecting seven items consistent with the design and objectives of the current study. An open-ended question was added to allow students to detail their personal experiences of doing homework with or without AI assistance. These provided more qualitative data and were used for within-data triangulation to contextualise, nuance or confirm the quantitative responses from the surveys and the AIDs. The data underwent a thematic analysis to reveal perspectives on the themes convergent with those targeted by the multiple choice questions from the survey and the topics of the AIDs. Due to the small size of the collected corpus (just over 1000 words), I performed the coding myself, iteratively, and extracted the relevant instances.

4. Results and Discussion

4.1 RQ1: How do Students Report Using GenAI Tools for Completing EAP Homework Assignments?

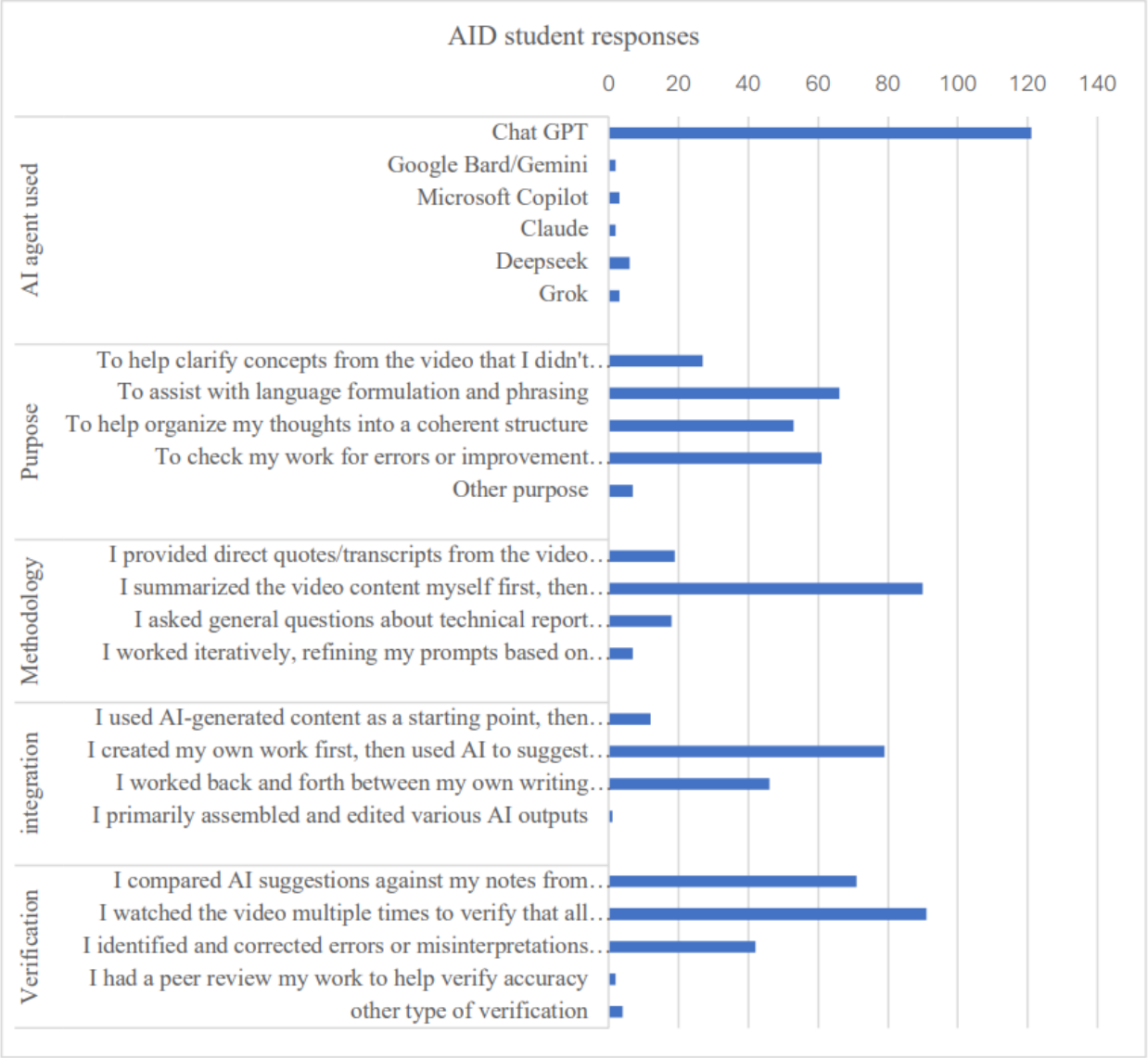


Figure 1. Students' responses from the AI Tools Usage Declaration (AID)

4.1.1 Choice of AI Tool

The findings from the AIDs are shown in Figure 1. A total of 137 (NT1=137) students used an AI tool to do their homework. By far the most used tool was ChatGPT (88.32%), confirming a recent study conducted by Yusuf et al. (2024), which showed that ChatGPT is the most widely utilised tool among both university lecturers and students. Six other different tools were used in much smaller proportions (e.g. DeepSeek=4.37%; Gemini, Claude=1.45%).

4.1.2 Purpose of AI Use

In terms of purpose, AI assisted mostly with language and formulating ideas more clearly or checking for errors and suggesting improvements. Students used AI tools to improve comprehension of video material as well as to organize their ideas and add coherence to their text. Other students mentioned assistance

with the layout and design of their output texts. These findings are consistent with reports such as Ngo and Hastie (2025), who show that students employ AI tools in EAP for a wide range of purposes, from those related to language mechanics to more EAP-specific ones, such as brainstorming and finding sources. Su et al. (2023) also found that collaborating with ChatGPT for writing increases the quality of ideas generated. My students displayed awareness of these capabilities and employed them in the process of solving the task, from the comprehension stage to the final decisions on the design and layout of the text they wrote. One student stated that 'doing my English homework was more manageable with AI assistance. I used it to help generate ideas, improve grammar, and structure my writing more clearly. It made the process faster and helped me learn by seeing better examples.'

The process of collaborating with the AI tools was broken down into three steps: methodology (how the AI tool was used), integration (how one's own work was integrated with AI-generated content and feedback) and verification (how accuracy and reliability of co-created content were ensured).

4.1.3 Methods of Engagement with AI Tools

The vast majority of students declared that they summarized the input material (video) themselves, seeking AI assistance to refine their work. A smaller number of students asked for support to understand the fundamental concepts of the lesson (technical report), and then they narrowed down the perspective they gained to the specific ideas from the video material provided by the teacher. In other words, students used AI to scaffold their comprehension of concepts before diving into more complex and specific aspects tackled in the EAP course. As one student put it: 'I enjoy doing homework with AI because it gives me the right information immediately instead of me searching for it; also, it can give me more specific explanations to help me understand better in all ways possible'.

Only a few students used an iterative process of evaluating AI and co-created work and refining prompts to get more specific results. Evaluation and iteration are crucial steps in working with GenAI tools to ensure accurate, specific and adapted output results (Schulhoff et al., 2025). This suggests that students may need training in order to be able to fully understand how GenAI works and can be used in most effective ways (Lee & Palmer, 2025), which supports the idea of integrating AI literacy and EAP in HE.

4.1.4 Integration of AI and Student-Generated Content

In terms of integration, students started from their own content and asked AI to suggest improvements or worked back and forth between their own writing and the AI suggestions, co-creating the final response. A small number of students used the AI-generated suggestions to build upon and refine for the final response.

4.1.5 Verification of Co-created Content

Fifty-seven students used multiple ways of verification to ensure the accuracy and reliability of co-created content. The most frequently used were watching the video multiple times to check own comprehension and fact-check AI summary, together with correction of own errors and those found in the AI generated content: 'I used AI to nitpick and make sure everything was all right.' and 'I used AI in correcting my grammar errors and to elevate my language by a slight margin.'

A few students appealed to a peer to check their work. One student reported that the only verification was asking the AI tool to check the final response for grammar or spelling errors. Integration and verification of co-created content are essential steps in completing EAP homework with the help of AI tools. The results offer only a glimpse of how students tackle them through self-reported practice, a limitation of the present study. However, they seem to indicate that students' engagement with the GenAI focuses less on their learning process and more on final output characteristics (e.g. style and punctuation). Leveraging GenAI potential for developing students' language learning process (e.g. iterative cycles of drafting that are critically evaluated) emerges as an important area of intervention when including AI literacy development in EAP curricula. Scaffolded and deliberate instruction on how to partner with AI tools throughout the learning process can empower students to become autonomous lifelong learners, fully aware of GenAI capabilities and limitations (Ngo & Hastie, 2025).

4.2 RQ2: What are the Students' Perceptions of the Benefits and Drawbacks of Using GenAI Tools for EAP Homework?

4.2.1 Perceived Benefits and Drawbacks of AI Tools Use

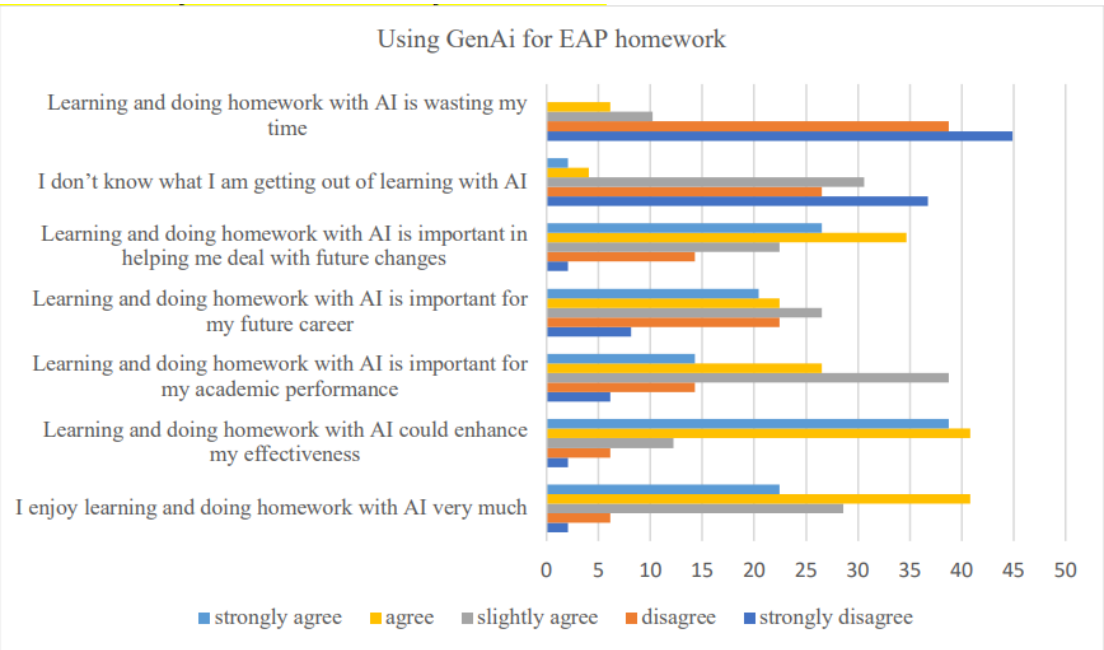


Figure 2. Students' perception of using GenAI for homework in EAP

Figure 2 summarizes the students' perceptions of the benefits and drawbacks of using GenAI for homework in EAP, based on the results of the post-activity survey. Only 49 students responded to this anonymous, non-mandatory questionnaire.

The survey results suggest a generally positive attitude toward learning and doing homework with AI. A clear majority of respondents expressed agreement or strong agreement with the statement that they enjoy using AI in this context, and similar support was evident for the view that AI could enhance their effectiveness. These responses indicate that many students not only recognise practical benefits but also experience genuine engagement and satisfaction when integrating AI into their academic work.

When asked about the importance of AI for their academic performance, future careers, and ability to adapt to future changes, the pattern remained favourable, albeit slightly more mixed. While a substantial proportion agreed or strongly agreed, a noticeable share of respondents chose "slightly agree" or lower levels of endorsement, suggesting that for some, the perceived value of AI is more tentative or conditional. This may reflect varying levels of familiarity, trust, or direct experience with AI's potential applications.

In contrast, negative perceptions—such as seeing AI as a waste of time or being unsure of its benefits—were less common but still present. Around one-third to nearly half of respondents disagreed or strongly disagreed with these negative statements, but a small percentage aligned with them. This minority may represent students who are either sceptical about AI's relevance or have not yet seen tangible benefits in their learning. Overall, the data points to a predominantly positive but not uncritical stance toward AI in academic contexts, with room for further confidence-building and practical demonstrations of its value.

4.2.2 Avoiding Engagement with AI Tools

This perspective on using GenAI for homework in EAP should be considered alongside data from the AIDs. Although only a minority (just over 14% of N = 160), some students reported not using any AI tool for their homework. Regardless of their reasons, this group preferred to work independently, avoiding AI assistance in their language learning process. Whether due to ethical concerns, confidence in their own language proficiency, technical difficulties, fear it would impede on their learning process ('I believe that, especially with writing in English, AI is very prone to replacing human input, and I try to develop this set of skills myself as much as possible'), or other factors, their motivation to abstain from AI use merits closer examination.

Existing research warns that the rapid proliferation of AI in education could deepen the digital divide (Beckman et al., 2025; Warschauer et al., 2023). Paywalls and other accessibility barriers can make AI tools unattainable for some students, preventing them from becoming as AI-savvy as their peers without such social, economic, or personal constraints.

In light of these concerns, higher education institutions bear a responsibility to mitigate inequities in access by ensuring that all students are provided with the requisite technology, training, and support. Integrating AI literacy into disciplinary curriculum, such as EAP courses, would not only address disparities in access but also equip students with the critical understanding and practical skills necessary to engage effectively with these rapidly evolving technologies.

5. Conclusion

This study set out to investigate how engineering students in a Romanian higher education institution use GenAI tools in the context of an EAP course, with the broader aim of assessing whether AI literacy should be embedded into the curriculum. Framed within an action research approach, the study not only documents student practices and perceptions but also serves as a structured reflection on the curricular conditions in which it takes place. The findings show that the vast majority of students who completed the AI Tools Declaration of Use had engaged with at least one AI tool—most often ChatGPT—primarily to refine language, clarify ideas, and ensure coherence in their work. Survey responses further revealed generally positive attitudes towards AI in learning, with students recognising its potential to enhance effectiveness, academic performance, and career preparedness. Nonetheless, a small but notable

minority preferred to work without AI assistance, highlighting diverse learner preferences and the need for a nuanced approach to AI integration.

The results also suggest that while students can employ AI effectively for language-related tasks, few demonstrated advanced practices such as iterative prompting or critical evaluation of AI output. This gap points to a lack of structured guidance and training, particularly in areas of responsible use, verification, and integration of AI-generated content. From an action research perspective, this insight functions as a call to refine both pedagogical strategies and curricular provisions to ensure that students not only access AI tools but also learn to use them critically and effectively. The presence of students who abstain from AI use—whether due to access, skills, or values—underscores the importance of equitable provision of both technology and AI literacy instruction, a responsibility that falls to higher education institutions. These results resonate with conclusions from Ngo and Hastie (2025) and Chan and Hu (2023), among others, who highlight that deliberate instruction should prepare all students to critically leverage GenAI potential for developing language skills and disciplinary communication while ethically acknowledging co-created content.

In keeping with the cyclical nature of action research, these findings will inform the next stage of my practice: adapting the EAP course to incorporate targeted AI literacy components that address the identified gaps while respecting diverse learner needs. This process of observing, analysing, and responding to classroom realities transforms the research into actionable change, aligning institutional goals with the evolving technological landscape (Vetter et al., 2024; Vettori & Warm, 2025). By systematically integrating these insights into future course iterations, the pedagogical and institutional context can evolve to support more equitable, informed, and effective engagement with GenAI in language learning.

While the findings provide valuable insights into students' engagement with GenAI tools, the study is limited by its reliance on self-reported data, which may not fully capture actual practices. The absence of observational or performance-based measures means that reported behaviours could not be independently verified. Future research could strengthen validity by incorporating task-based assessments or classroom observations alongside surveys, providing a more comprehensive picture of AI use in EAP contexts.

References

Banegas, D. L., & Consoli, S. (2020). Action research in language education. In J. McKinley & H. Rose (Eds.), *The Routledge Handbook of Research Methods in Applied Linguistics* (pp. 176–187). Routledge.

Beckman, K., Apps, T., Howard, S. K., Rogerson, C., Rogerson, A., & Tondeur, J. (2025). The GenAI divide among university students: A call for action. *The Internet and Higher Education*, 101036.

Chan, C. K. Y., & Hu, W. (2023). Students' voices on generative AI: perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education*, 20(1). https://doi.org/10.1186/s41239-023-00411-8

Chiu, T. K. F., Xia, Q., Zhou, X., Chai, C. S., & Cheng, M. (2023). Systematic literature review on opportunities, challenges, and future research recommendations of artificial intelligence in education. *Computers and Education: Artificial Intelligence*, 4, Article 100118. https://doi.org/10.1016/j.caeai.2022.100118

Darvin, R. (2025). Identity and investment in the age of generative AI. *Annual Review of Applied Linguistics*, 45, 10-27. https://doi.org/10.1017/S0267190525100135

Godwin-Jones, R. (2022). Partnering with AI: Intelligent writing assistance and instructed language learning. *Language Learning & Technology*, 26(2), 5-24. http://doi.org/10125/73474

Ingley, S. J., & Pack, A. (2023). Leveraging AI tools to develop the writer rather than the writing. *Trends in Ecology and Evolution*, 38(9), 785–787. https://doi.org/10.1016/j.tree.2023.05.007

Khodi, A., & Curle, S. (2025, July 30). Can AI Replace Human Insight? Evaluating the Quality of Feedback from Human Tutors Versus ChatGPT on EFL Students' Argumentative Writing [Conference presentation]. *International Conference on Globalisation/Deglobalisation in Languages, Education, Culture and Communication*, Manchester, UK.

Lee, D., Arnold, M., Srivastava, A., Plastow, K., Strelan, P., Ploeckl, F., Lekkas, D., & Palmer, E. (2024). The impact of generative AI on higher education learning and teaching: A study of educators' perspectives. *Computers and Education: Artificial Intelligence*, 6, Article 100221. https://doi.org/10.1016/j.caeai.2024.100221

Lee, D., & Palmer, E. (2025). Prompt engineering in higher education: a systematic review to help inform curricula. *International Journal of Educational Technology in Higher Education*, 22(1). https://doi.org/10.1186/s41239-025-00503-7

Li, J., King, R. B., Chai, C. S., Zhai, X., & Lee, V. W. Y. (2025). The AI Motivation Scale (AIMS): a self-determination theory perspective. *Journal of Research on Technology in Education*, 1-22. https://doi.org/10.1080/15391523.2025.2478424

Mahapatra, S. (2024). Impact of ChatGPT on ESL students' academic writing skills: a mixed methods intervention study. *Smart Learning Environments*, 11(1). https://doi.org/10.1186/s40561-024-00295-9

Nelson, A. S., Santamaría, P. V., Javens, J. S., & Ricaurte, M. (2025). Students' perceptions of generative artificial intelligence (GenAI) use in academic writing in English as a foreign language. *Education Sciences*, 15(5). https://doi.org/10.3390/educsci15050611

Ngo, T. N., & Hastie, D. (2025). Artificial Intelligence for Academic Purposes (AIAP): Integrating AI literacy into an EAP module. *English for Specific Purposes*, 77, 20–38. https://doi.org/10.1016/j.esp.2024.09.001

Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., Liu, A., Si, C., Li, Y., Gupta, A., Han, H., Schulhoff, S., Dulepet, P. S., Vidyadhara, S., Ki, D., Agrawal, S., Pham, C., Kroiz, G., Li, F., Tao, H., Srivastava, A., & Resnik, P. (2025). *The Prompt Report: A Systematic Survey of Prompt Engineering Techniques. Cornell University*. http://arxiv.org/abs/2406.06608

Southworth, J., Migliaccio, K., Glover, J., Glover, J. N., Reed, D., McCarty, C., Brendemuhl, J., & Thomas, A. (2023). Developing a model for AI Across the curriculum: Transforming the higher education landscape via innovation in AI literacy. *Computers and Education: Artificial Intelligence*, 4, Article 100127. https://doi.org/10.1016/j.caeai.2023.100127

Su, Y., Lin, Y., & Lai, C. (2023). Collaborating with ChatGPT in argumentative writing classrooms. *Assessing Writing*, 57, Article 100752. https://doi.org/10.1016/j.asw.2023.100752

Vetter, M. A., Lucia, B., Jiang, J., & Othman, M. (2024). Towards a framework for local interrogation of AI ethics: A case study on text generators, academic integrity, and composing with ChatGPT. *Computers and Composition*, 71, Article 102831. https://doi.org/10.1016/j.compcom.2024.102831

Vettori, O., & Warm, J. (2025). The race for AI skills as an obstacle course: Institutional challenges and low threshold suggestions. *Project Leadership and Society*, 6, Article 100183. https://doi.org/10.1016/j.plas.2025.100183

Warschauer, M., Tseng, W., Yim, S., Webster, T., Jacob, S., Du, Q., & Tate, T. (2023). The affordances and contradictions of AI-generated text for writers of english as a second or foreign language. *Journal of Second Language Writing*, 62, Article 101071. https://doi.org/10.1016/j.jslw.2023.101071

Yang, Y., Zhang, Y., Sun, D., He, W., & Wei, Y. (2025). Navigating the landscape of AI literacy education: insights from a decade of research (2014–2024). *Humanities and Social Sciences Communications*, 12(1), Article 374. https://doi.org/10.1057/s41599-025-04583-8

Yusuf, A., Pervin, N., & Román-González, M. (2024). Generative AI and the future of higher education: a threat to academic integrity or reformation? Evidence from multicultural perspectives. *International Journal of Educational Technology in Higher Education*, 21(1), Article 21. https://doi.org/10.1186/s41239-024-00453-6

Sonia C. Munteanu, PhD (Sonia.Munteanu@lang.utcluj.ro) is an Associate Professor at the Technical University of Cluj-Napoca (TUC-N), Romania, where she teaches English for specific purposes, academic English, intercultural communication, and Romanian as a foreign language. Her main research interests are language teaching and learning, applied linguistics, English medium education, intercultural communication, and internationalization of higher education.

Annex: AI Tools Usage Declaration

Student Information

Name: _____

Group: _____

Declaration

I hereby declare the following regarding my use of AI-assisted tools in completing this assignment:

- [ ] I have NOT used any AI-assisted tools in completing this assignment.
- [ ] I HAVE used AI-assisted tools in completing this assignment, and I provide details below:

1. Tools used (select all that apply):

- [ ] ChatGPT
- [ ] Google Bard/Gemini
- [ ] Microsoft Copilot
- [ ] Claude
- [ ] Other (please specify): _____

2. Purpose (select all that apply):

- [ ] To help clarify concepts from the video that I didn't understand
- [ ] To assist with language formulation and phrasing
- [ ] To help organize my thoughts into a coherent structure
- [ ] To check my work for errors or improvement suggestions
- [ ] Other (please specify): _____

3. Methodology (select one):

- [ ] I provided direct quotes/transcripts from the video and asked AI to summarize them
- [ ] I summarized the video content myself first, then asked AI to help refine my work
- [ ] I asked general questions about technical report writing and integrated the AI's responses with the video content
- [ ] I worked iteratively, refining my prompts based on initial AI responses

4. Integration (select one):

- [ ] I used AI-generated content as a starting point, then substantially revised and added my own contribution
- [ ] I created my own work first, then used AI to suggest improvements or alternative phrasings
- [ ] I worked back and forth between my own writing and AI suggestions, creating an integrated final product
- [ ] I primarily assembled and edited various AI outputs

5. Verification (select all that apply):

- [ ] I compared AI suggestions against my notes from the video to ensure accuracy
- [ ] I watched the video multiple times to verify that all key points were included correctly
- [ ] I identified and corrected errors or misinterpretations in the AI-generated content
- [ ] I had a peer review my work to help verify accuracy
- [ ] Other (please specify): _____

Research Papers in Language Teaching and Learning
Vol. 16, No. 1, March 2026, 60-78
ISSN: 1792-1244
Available online at http://rpltl.eap.gr

Creative and Critical Integration of Artificial Intelligence in EFL Learning

Sophia Kouzouli

The increasing integration of artificial intelligence (AI) in education is reshaping language teaching and creating new conditions for pedagogical innovation while also raising important ethical and educational questions. This paper presents a values-based lesson design for the creative and critical integration of artificial intelligence in English as a Foreign Language education, illustrated through the design and classroom implementation of the lesson "Odyssey 2.0: Values Recharged." AI is conceptualized as a learning resource that supports linguistic development, critical digital literacy, creativity, and ethical awareness through constructivist, experiential, inquiry-based, and reflective learning practices. The lesson design integrates multimodal activities, including digital storytelling, collaborative inquiry, role-play, debate, and creative production, within a learner-centred, values-based educational focus. The paper discusses implications for curriculum design, teacher practice, and formative assessment, while addressing ethical challenges related to bias, dependency, and data protection. Overall, it illustrates how purposeful AI integration, guided by clear pedagogical principles, can enrich EFL learning and support the development of critically aware and socially responsible learners.

Keywords: Artificial Intelligence Pedagogy; EFL Pedagogy; Critical Digital Literacy; Values-Based Learning.

---

1. Introduction

The integration of AI into educational practice constitutes one of the most transformative developments in contemporary pedagogy, as the OECD (2026) highlights the potential of generative AI to transform the quality and effectiveness of learning and education systems. In EFL contexts, Selwyn (2019) calls for critical reflection on how AI-supported technologies reshape educational roles and practices. More recent research shows that such technologies increasingly mediate lesson design, assessment practices, learner engagement, and teacher

roles, contributing to the emergence of dynamic, multimodal language learning environments (Frøsig & Romero, 2024). UNESCO (2021) likewise argues that these developments require sustained reflection and ethical awareness.

From a sociocultural perspective, learning is understood as a socially mediated process in which knowledge emerges through interaction, dialogue, and participation in meaningful activity (Vygotsky, 1978). Within this view, learning is not conceived as the mere acquisition of linguistic forms but as the development of communicative competence through engagement with cultural, social, and cognitive practices. Constructivist learning theory (Bruner, 1996) and experiential learning theory (Kolb, 1984) further emphasize the learner's active role in developing understanding through experience, inquiry, reflection, and collaboration.

Building on these theoretical foundations, the article presents and analyses a pedagogical lesson design that illustrates how artificial intelligence can be meaningfully integrated into EFL learning within a human-centred and values-oriented educational framework. In this context, the paper explores two key pedagogical issues: first, how artificial intelligence can be used in EFL teaching in ways that support linguistic development and creative language use; and, second, how AI-supported learning activities can foster critical digital literacy and reflective engagement with technology.

2. Theoretical Framework

This section draws on insights from contemporary theories of learning, language education, and technology integration, incorporating constructivist and sociocultural perspectives as well as research on AI in language education, critical digital literacy, and values-based pedagogy. From a constructivist and sociocultural perspective, learning is understood as an active, socially situated process in which knowledge is constructed through interaction, dialogue, and participation in meaningful activity (Dewey, 1938; Vygotsky, 1978). From this perspective, communicative competence emerges through learners' engagement in cognitive, social, and cultural practices rather than through the accumulation of isolated linguistic forms. Experiential learning theory emphasizes the process through which understanding is developed via experience, reflection, conceptualization, and experimentation (Kolb, 1984), while inquiry-based learning positions learners as active investigators who construct knowledge through problem-solving and exploration (Hmelo-Silver, 2004). Collaborative learning theories highlight the importance of interaction, shared responsibility, and social negotiation in knowledge construction, principles that underpin effective communicative language teaching (Johnson & Johnson, 2009).

Concerns that AI-supported assistance may lead to cognitive offloading have been raised in recent discussions; however, this position has been challenged in subsequent theoretical and critical work. Potkalitsky (2025) argues that assistance has always been integral to human learning and knowledge construction rather than an external threat to cognition. From this perspective, tools and technologies function as extensions of human thinking within socially and culturally mediated learning processes. This view supports an understanding of AI not as a replacement for human cognition, but as a pedagogical resource whose educational value depends on intentional and reflective instructional use.

The development of critical digital literacy is a central requirement of contemporary language education. Critical digital literacy refers to learners' ability to critically access, interpret, evaluate, and create digital information while recognising the social, ethical, and political implications of digital technologies and algorithmic systems (Gee, 2015; European Commission, 2026). The expansion of digital technologies has contributed to a broader understanding of literacy as a constellation of digital, media, and critical competences (Gee, 2015; Kern, 2015). In AI-enhanced learning environments, learners are required to analyse, evaluate, and create digital content while understanding the ethical, social, and cultural implications of algorithmic systems. Such engagement enables learners to recognize bias and limitations in AI-generated content and to participate actively in knowledge construction (European Commission, 2026).

Recent discussions on artificial intelligence in education and the development of critical digital literacy increasingly emphasise the importance of maintaining a pedagogically guided AI use. Within this perspective, human-centred artificial intelligence in education refers to the development and use of AI systems that support human learning, strengthen teacher professional judgement, and promote learner agency, well-being, and authentic participation in the learning process (Holmes et al., 2019; UNESCO, 2023). Human-centred pedagogy, as reported by Anastasiades et al. (2024), places learners' cognitive, emotional, social, and ethical development at the core of educational practice, prioritizing relationships, inclusivity, and personal growth over technological efficiency. Lazcano-Quintana (2024) argues that values-based education reinforces this orientation by embedding moral, cultural, and civic values into the learning process, fostering learners' understanding of responsibility, perseverance, cooperation, respect, and empathy alongside academic development.

The effective use of artificial intelligence in education requires its contextualised alignment with pedagogical aims and instructional design. In this context, the pedagogical integration of AI can be understood as the intentional alignment of AI tools with learning objectives, instructional strategies, and assessment practices in ways that support meaningful learning processes and sustained cognitive engagement (Mishra & Koehler, 2006; OECD, 2026). The pedagogical integration of AI is further informed by the Technological Pedagogical Content Knowledge (TPACK) model, which conceptualizes effective teaching as the intersection of content knowledge, pedagogical knowledge, and technological knowledge (Mishra & Koehler, 2006). Complementary models such as SAMR (Substitution, Augmentation, Modification, Redefinition) inform the progressive transformation of learning tasks through technology, guiding practice from substitution toward pedagogical redefinition (Puentedura, 2013). Research on pedagogically mediated AI use reinforces the importance of aligning technological innovation with ethical responsibility, transparency, and learner well-being, as emphasised by Holmes et al. (2019) and Luckin (2018).

3. The "Odyssey 2.0: Values Recharged" Pedagogical Lesson Design

The article presents a pedagogical lesson design that illustrates the meaningful use of AI within EFL teaching in a Greek secondary school context, addressing 12-13 years oldd learners whose language proficiency ranges from A2 to B1. The lesson was implemented as a model pedagogical practice intended to illustrate how AI-supported activities can be integrated into EFL teaching. It was developed as a 90-minute teaching sequence, offering a concrete pedagogical illustration of the lesson design in authentic classroom conditions. The lesson was

also designed within the framework of the B2 level of ICT teacher training in Greece, ensuring systematic alignment among pedagogical objectives, technological integration, and curriculum requirements.

The lesson design presupposes a set of linguistic, digital, and cultural prerequisites that support learners' engagement with the tasks. At the linguistic level, learners are expected to have a basic command of English, including familiarity with vocabulary related to the description of people and values, the ability to comprehend and produce simple narrative and descriptive texts, and the capacity to participate in basic dialogue and express opinions. In terms of digital competence, learners are expected to be familiar with the basic use of common digital tools such as Canva, Lumen5, Quizlet, ChatGPT, and Suno, as well as with searching for information online and interpreting digital texts and images. In addition, learners are expected to demonstrate cultural and cognitive readiness, including familiarity with the mythology and historical background of Odysseus, an emerging capacity for critical thinking and for comparing past and present perspectives, and an initial awareness of the role of values in contemporary society.

Language development is set in a meaningful cultural and ethical contextthat adheres to the principles of the CEFR (Council of Europe, 2020), enabling learners to explore identity, heroism, responsibility, and social values while simultaneously strengthening communicative competence, in line with constructivist and sociocultural views of learning. Drawing inspiration from classical narratives of heroism and ethical development, the lesson connects cultural heritage with contemporary social realities through activities such as the exploration of Odysseus' return, the comparison of mythological and modern forms of heroism, and the examination of everyday heroes in local and social contexts. It is pedagogically aligned with thematic units of the official EFL curriculum related to heroes, values, descriptions of people and events, and narrative texts, while also functioning interdisciplinarily through connections with the subjects of Ancient Greek (the Odyssey), Literature, History, Art, and ICT.

At the core of the lesson design lies the principle of pedagogical integration, understood as the alignment of technological tools with clearly articulated learning objectives, instructional methods, assessment practices, and educational values. AI is not introduced as a technological novelty but is integrated as a mediational pedagogical resource that supports cognitive engagement, scaffolds linguistic production, and enhances creativity, differentiation, and inclusion, consistent with the learner-centred AI use in education articulated by OECD (2026). Research on the impact of AI technologies on student learning and performance further underscores the need for balanced integration that prioritises pedagogical values and preserves critical thinking and learner agency, as demonstrated by Vieriu & Petrea (2025) and emphasised in systematic reviews of human-centred AI design by Schmager et al. (2025). AI-supported activities are embedded within coherent instructional sequences that encourage learners to analyse, compare, and generate content, to connect traditional and contemporary conceptions of heroism, and to revise linguistic output through collaborative reflection.

The lesson integrates a wide range of learning activities, including inquiry tasks, language practice, role-play, debate, multimodal text production, digital storytelling, and guided use of AI tools. These activities are intentionally combined to foster active participation, creative expression, and higher-order thinking, while supporting meaningful communication and learner agency throughout the instructional process as suggested by Godwin-Jones (2024).

It is structured around three interrelated dimensions that support holistic learner development. Language development is fostered through authentic communicative practices involving narrative construction, lexical expansion, dialogue, and multimodal expression, supported by AI-enabled resources that provide rich linguistic input and scaffolded practice across multiple semiotic modes (Warschauer & Kern, 2000). Critical and digital competence is cultivated as learners analyse, evaluate, and co-create content using AI tools, developing informed awareness of the affordances and limitations of intelligent systems and strengthening responsible digital agency (Knobel & Lankshear, 2006). For example, during the role-play and debate stations, students use AI-generated suggestions from ChatGPT but are required to evaluate their accuracy and relevance before incorporating them into their arguments, thereby practising critical judgement in AI-supported environments (European Commission, 2026; OECD, 2026). Values and social awareness are integrated through ethical reflection, social responsibility, intercultural understanding, and the exploration of core values such as perseverance, loyalty, empathy, courage, cooperation, and respect, supported by storytelling, role-play, collaborative projects, and reflective dialogue.

The learning environment fostered by the Odyssey 2.0 lesson design is intentionally interactive and learner-centred. Students assume active roles in the construction of knowledge through collaborative inquiry, creative production, debate, and reflection, while the teacher functions as a facilitator who designs learning experiences, guides reflection, and supports learners' cognitive, emotional, and social development (Schön, 1983; Nicol & Macfarlane-Dick, 2006).

4. Learning Objectives and Expected Competence Development

Learning objectives are formulated as expected outcomes that guide learners' engagement with language skills, technology, and values in a coherent and interrelated manner. Therefore, the expected learning outcomes encompass linguistic, digital, cognitive, social, and ethical competences, following an integrated developmental process that can support both academic achievement and the formation of socially responsible individuals.

With regard to linguistic development, learners are expected to enhance both receptive and productive skills in English through engagement with spoken and written texts, including textual and audiovisual material. They are encouraged to comprehend oral and written discourse and to produce simple spoken and written texts using appropriate vocabulary and basic structures. In particular, learners are facilitated to generate short narratives and descriptions related to heroes and values, to transfer simple information from source texts into creative products (such as posters or short digital texts), and to participate in dialogues and role-play activities, expressing ideas and arguments in English at an appropriate level. Through these communicative practices, learners are supported to strengthen fluency, accuracy, and confidence in using English for meaningful interaction, as research on communicative language teaching demonstrates (Warschauer & Kern, 2000).

In relation to technology use and digital literacy, learners are expected to develop the ability to use AI-supported tools for the creation of multimodal texts. They are also assisted to analyse and evaluate content generated with the support of AI and to develop an initial understanding of both the strengths and the limitations of AI in educational and communicative contexts. Furthermore, learners are empowered to collaborate in digital

environments, producing content through group work and shared digital practices, thereby strengthening both technical and collaborative skills.

Cognitive and learning-related outcomes focus on the development of critical thinking and reflective engagement with content. Learners are expected to analyse and compare past and present perspectives, particularly in relation to concepts of heroism and values, and to connect historical and mythological references with contemporary social contexts. Through inquiry-based and creative activities, learners are expected to cultivate problem-solving abilities, to reflect on their learning process, and to develop creativity through interdisciplinary tasks that combine language, culture, and technology.

Social, ethical, and values-based outcomes constitute a central dimension of the lesson. Learners are expected to recognize the importance of social responsibility and ethical awareness in the contemporary world, to reflect on values such as cooperation, responsibility, and respect, and to develop a more informed and balanced stance toward technology. In addition, learners are expected to enhance intercultural awareness and to appreciate the value of cultural heritage, while strengthening their ability to communicate effectively and collaborate with others. Through these experiences, the lesson aims to support learners not only as developing users of English but also as active, reflective, and socially aware participants in an interconnected world.

5. Implementation of the Odyssey 2.0 Lesson: Learning Activities and Tools

The lesson was implemented in a Greek public lower secondary school (Gymnasium). The class consisted of sixteen first-grade Gymnasium students (approximately 12–13 years old) with an overall English proficiency level corresponding to A2 of the Common European Framework of Reference for Languages. It was delivered in an authentic classroom environment as a model lesson by the Education Advisor for EFL teachers and was attended by English language teachers. It lasted for two teaching periods, ninety minutes. The student samples included in the appendices derive from this classroom implementation of the model lesson. It was implemented as a coherent learner-centred sequence that integrates inquiry, collaboration, creative production, and reflection within a technologically enriched environment. The activities were designed to encourage active participation, communication, and reflection, while ensuring that AI tools functioned as pedagogical support for classroom interaction and discussion.

The overall sequencing of activities supported gradual conceptual development, moving from the activation of prior knowledge to collaborative exploration, creative expression, and reflective evaluation. This structure ensured alignment between learning objectives, classroom practices, and assessment, and prepared learners to engage productively with both linguistic content and digital tools.

The lesson was organized into two teaching periods. The first teaching hour introduced learners to the central thematic focus of heroism and values through guided exploration and language-focused activities. Activity 1, “The Return of Odysseus”, activated prior knowledge through brief brainstorming on the concept of a hero, followed by a guided discussion after viewing a short film excerpt depicting Odysseus’ return and his recognition by his dog, Argos. This activity supported comprehension, thematic engagement, and initial identification of values reflected in the narrative.

In Activity 2, “The Journey Timeline”, students worked collaboratively in small groups to reconstruct key events of Odysseus’ journey. Learners worked with short narrative texts adapted from a child-friendly online encyclopedic resource (Kiddle), which were modified to match their language proficiency level using an AI-supported text rewriting tool (Text Rewriter, Magic School; see Appendix A). The adapted texts were presented through the digital tool Canva, and additionally shared as printed cards, and read collaboratively by the groups. Through comparing information across texts and sequencing episodes chronologically, they practiced descriptive language, developed narrative coherence, and engaged in oral interaction in a cooperative learning context.

In Activity 3, “Qualities of a Hero”, learners explored lexical items related to personal qualities and values through guided discussion and structured vocabulary practice, supported by digital flashcards created in Quizlet, which strengthened their understanding of key terms required for subsequent tasks. This activity prepared the ground for Activity 4, “Values Map”, during which learners participated in a digital brainstorming activity using AnswerGarden, contributing words that described the qualities of a hero. As learners submitted their responses, a shared visual representation of ideas was gradually formed, allowing the group to observe recurring values and patterns. Through this process, learners engaged in collaborative reflection on contemporary forms of heroism, enhanced participation and cooperation, and became familiar with the use of digital tools for collective idea generation.

In Activity 5, “Contemporary Heroes in the Neighbourhood”, learners engaged in guided brainstorming on examples of everyday heroism, focusing on figures such as volunteers, firefighters, and activists. This initial phase aimed to encourage reflection on contemporary forms of heroism and to connect abstract values with real-life social contexts. Learners then read a short informational text related to local or community-based heroes, which had been generated using an AI-supported tool (Informational Text, Magic School; see Appendix B). Working in groups, they identified and recorded examples of contemporary heroes, thereby reinforcing reading comprehension, collaborative interaction, and values-based reflection.

The second teaching hour was organized around a learning stations model, enabling participants to engage with the lesson theme through multiple forms of expression and interaction. The stations collectively addressed storytelling, role-play, debate, music creation, and visual representation, offering diverse pathways for exploring contemporary heroism and associated values. This structure supported differentiated participation, collaborative learning, and active language use, while integrating digital and AI-supported tools in meaningful communicative tasks.

The first station, the Storytelling Corner, focused on narrative creation and creative language use. Learners worked collaboratively to produce a short story about a contemporary hero using an AI-supported storytelling tool (Storybird). Examples of student outputs produced during the learning stations are presented in Appendix C. They were guided to consider basic narrative elements, such as character, setting, action, and values, and to incorporate previously introduced vocabulary related to personal qualities. Visual prompts supported idea development, while teacher mediation ensured attention to coherence, clarity, and appropriate language use. At the conclusion of the station work, learners engaged in structured self-reflection and self-assessment using a reflective rubric. The self-assessment focused on language use, vocabulary application, organisation of ideas, collaboration, and the

purposeful use of AI tools during the storytelling process. This final stage encouraged learners to reflect on their learning experience, recognize areas of progress, and identify aspects requiring further improvement, reinforcing metacognitive awareness and responsible engagement with digital technologies.

In the second learning station, Digital Creators, learners worked collaboratively to design and present a short video portraying a contemporary hero. Using the AI-supported video creation tool Lumen5, students transformed ideas into multimodal products by combining images, short texts, and basic narration. Lexical resources previously practiced through Quizlet were deliberately incorporated, ensuring that language use remained aligned with the lesson's linguistic objectives while supporting creative expression.

This station promoted language practice, creative writing, and digital literacy through purposeful technology use. Learners planned content, selected visual elements, and presented their ideas orally, strengthening narrative coherence and collaborative communication, while engaging in brief guided self-reflection on their language use, collaboration, and use of digital tools. Teacher guidance ensured that AI functioned as a supportive tool rather than a substitute for learner agency, maintaining pedagogical coherence and meaningful engagement with both language and values.

In the third learning station, learners engaged in short role-play activities centred on the theme of everyday heroism. Working collaboratively, students created and performed brief scenes in which a hero helps someone in danger, using three to four spoken sentences per role. A chatbot tool (ChatGPT) was used selectively to support idea generation and scenario development, while the primary emphasis remained on oral interaction, communicative clarity, and the meaningful use of English in context. Following the role-play, learners completed a structured self-assessment rubric focusing on language use, contribution to group work, use of the AI tool, and engagement during the activity. This reflective process supported learners' awareness of their communicative performance, collaboration, and learning strategies, reinforcing formative assessment and metacognitive reflection as integral components of the learning process.

In the fourth learning station, learners participated in a structured debate focusing on the question of heroism in contemporary society. Students were divided into two groups and discussed whether Odysseus or a modern local hero is more relevant today, drawing on values, social contribution, and their role in contemporary society. A chatbot tool (ChatGPT) was used selectively by each group to support the formulation of one argument, while the debate emphasized spoken interaction, justification of opinions, and respectful dialogue in English.

Following the debate, learners completed a self-assessment rubric addressing their participation in discussion, quality of argumentation, use of the AI tool, collaboration within the group, and respect for differing viewpoints. This reflective activity supported learners' awareness of their communicative strategies, critical thinking, and interactional skills, reinforcing formative assessment and reflective learning as integral elements of the instructional process. In the end, learners engaged in guided reflection through an exit ticket activity. They were invited to articulate one new insight gained about heroes or values, identify the learning station they found most engaging, and consider how one of the values discussed could be applied in their everyday lives. This reflective process supported personal

meaning-making, encouraged transfer of learning beyond the classroom, and fostered awareness of values such as perseverance, courage, and responsibility.

Assessment was implemented through a combination of systematic observation, analytic rubrics, and a short digital quiz. Rubrics were used both for self-assessment and teacher assessment, focusing on language use, collaboration, engagement with AI tools, and quality of produced work. A sample of the student self-assessment rubric and reflective feedback is provided in Appendix D. This multimodal approach emphasized formative feedback and learner reflection, ensuring alignment among learning objectives, learning activities, and evaluation practices.

6. Role of Artificial Intelligence in Language Learning

Artificial intelligence is used as a mediating pedagogical resource that supports language development, creative expression, and critical engagement. Its role is explicitly framed within a human-centred educational logic, in which pedagogical intentionality, ethical awareness, and learner agency guide all technological use, as emphasized in UNESCO's (2023) guidance on learner-centred approaches to generative AI in education.

At the linguistic level, AI-supported tools function as scaffolding mechanisms that assist learners in exploring vocabulary, sentence structure, and stylistic alternatives. They facilitate experimentation, revision, and reflection, allowing learners to test hypotheses about language use and refine their output through guided interaction. Teacher mediation remains central in this process, as educators support learners in evaluating AI-generated suggestions, making contextually appropriate choices, and developing pragmatic awareness. In this way, AI supports linguistic reflection rather than replacing cognitive effort or pedagogical judgment (Warschauer & Kern, 2000).

Beyond linguistic support, AI can contribute to multimodal meaning-making. Multimodal meaning-making, as Kress (2010) demonstrates in his theory of multimodal communication, refers to the construction and communication of meaning through the interaction of multiple semiotic modes, such as language, image, sound, and visual design. Through the integration of text, image, sound, and visual design, learners engage in creative composition that accommodates diverse learning styles and expressive preferences. For instance, in the "Digital Creators" station, students produce short videos portraying contemporary heroes using Lumen5, combining images, captions, and narration to communicate values such as courage or cooperation through multimodal storytelling (Kress, 2010; OECD, 2026). This multimodal engagement strengthens learner motivation and confidence, encouraging productive risk-taking and experimentation, as research on digital language learning environments has shown (Warschauer & Kern, 2000), while maintaining alignment with curricular goals.

A central pedagogical objective of the Odyssey 2.0 model is the cultivation of critical engagement with AI-generated content. Learners are encouraged to question accuracy, relevance, and potential bias, developing awareness of the limitations of algorithmic systems. Ethical considerations, such as authorship, are embedded within instructional practice, ensuring that technological engagement remains aligned with educational values and social responsibility.

Finally, AI contributes to a reconfiguration of classroom roles. Learners assume increased responsibility for decision-making, collaboration, and knowledge construction, while teachers function as learning designers and mentors who facilitate inquiry and reflection. Thus, AI amplifies human interaction, creativity, and pedagogical purpose.

7. Discussion: Pedagogical Value and Design Insights

The Odyssey 2.0: Values Recharged lesson design illustrates one possible approach to integrating AI in ways that may enrich EFL pedagogy within a human-centred and values-oriented pedagogical perspective. AI serves as a pedagogical resource that supports clearly articulated pedagogical intent, learning objectives, and ethical considerations, aligning technology use with broader educational purposes (Biesta, 2015). To support dissemination and pedagogical reuse, the learning design has also been represented as a Learning Activity Management System (LAMS) sequence and shared through the LAMS Community repository.

A key pedagogical feature of the lesson is the attempt to position the language classroom as an active space for inquiry, collaboration, and creative production. Through multimodal tasks, collaborative problem-solving, and reflective dialogue, learners engage with language in ways that encourage participation and meaning-making. Research suggests that such participatory learning environments have been shown to enhance learner motivation and cognitive engagement (Hattie, 2009; Voogt et al., 2013). In this lesson, the learning-stations model was used to organise activities such as multimodal storytelling, collaborative discussion, and digital content creation, allowing learners to approach language use through multiple modes and perspectives.

Within this design, AI tools can function as mediating resources that provide linguistic suggestions, creative prompts, and opportunities for multimodal expression. At the same time, teacher mediation remains central in supporting interpretation, revision, and critical evaluation of AI-generated content. Learners are encouraged to examine AI-generated outputs, assess their clarity and appropriateness, and reflect on their limitations, thereby maintaining an active and reflective stance toward technology use. They should consider their clarity and relevance, and reflect on their potential limitations. In this sense, the design aims to provide insights for the pedagogically guided use of AI in classroom practice and to position it as a resource that can stimulate questioning, evaluation, and discussion.

The lesson design also highlights potential shifts in classroom roles that are increasingly discussed in the literature on technology-supported learning. Teachers assume the role of learning designers and facilitators who orchestrate inquiry, support reflection, and maintain pedagogical coherence, while learners take greater responsibility for decision-making, collaboration, and knowledge construction. These dynamics resonate with discussions of classroom agency in technology-mediated learning environments, as described by Vedder-Weiss (2025). Such a redistribution of agency may contribute to the development of learning environments in which technology supports interaction, creativity, and ethical awareness, as emphasised by Holmes et al. (2019) in their discussion of pedagogically guided uses of AI in education.

Additionally, this example connects with emerging perspectives that encourage educational practices to engage constructively with the presence of AI in learning environments. As Fitzpatrick (2026) argues, educational engagement with AI should be grounded in

fundamental pedagogical questions concerning educational purpose and value. By embedding AI use within narratives of cultural heritage, social relevance, and ethical inquiry, the Odyssey 2.0 lesson design illustrates how language learning activities can integrate technological innovation while remaining attentive to human values and pedagogical responsibility.

Finally, it is important to acknowledge that the lesson represents a single classroom implementation and is presented primarily as a pedagogical design example. Further work could explore how similar approaches may be adapted across different educational contexts and examine more systematically how teachers and learners experience AI-supported language learning activities

8. Ethical Issues and Limitations

While the integration of AI in the Odyssey 2.0 lesson design offers significant pedagogical benefits, it also raises ethical and practical considerations that require careful attention. Recent international policy work has likewise emphasized the importance of responsible, learner-focused approaches to AI in education, highlighting issues of equity, governance, transparency, and learner protection as central to sustainable educational innovation (OECD, 2023). Facer & Selwyn (2021) advocate that a human-centred approach to AI-enhanced learning necessitates ongoing reflection on issues related to learner protection, equity, and responsible technology use.

One central concern involves data privacy and learner safety. AI-based platforms may collect and process user data, making informed tool selection and responsible classroom use essential. In the context of the present lesson, all digital tools were accessed through school-created accounts set up and managed in the school computer laboratory, ensuring controlled use and institutional oversight. This practice reflects institutional digital policy and aligns with a TPACK-informed approach, as well as with the principles of B2-level ICT teacher training, which emphasize pedagogically grounded, ethically responsible, and context-aware technology integration. Educators, therefore, need to remain attentive to issues of consent, transparency, and data governance, particularly in school contexts involving minors, as highlighted in discussions of digital education governance and data-driven policy instruments (Ben Williamson, 2016).

Algorithmic bias represents a further limitation. AI-generated content may reproduce cultural, social, or linguistic biases embedded in training data, potentially reinforcing narrow representations or dominant norms. The Odyssey 2.0 lesson design addresses this risk by embedding critical digital literacy into practice, encouraging learners to question AI outputs and recognize their limitations (Gee, 2015).

Finally, the potential for overreliance on AI tools must be acknowledged. Excessive dependence may reduce opportunities for sustained cognitive engagement or independent problem-solving, and systematic evidence indicates that over-reliance on AI dialogue systems can negatively affect students' decision-making, critical thinking, and analytical reasoning. Zhai et al. (2024), therefore, emphasize balanced and reflective integration, positioning AI as a scaffold that supports learning without displacing creativity, critical thinking, or human agency.

9. Implications for EFL Teaching and Conclusion

The lesson design presented in this paper illustrates how artificial intelligence can be pedagogically integrated into EFL teaching in ways that support linguistic development, critical digital literacy, and values-based reflection and align technological innovation with curricular coherence and broader educational aims. Its classroom implementation highlights several implications for contemporary language education.

First, EFL curriculum design can benefit from moving beyond isolated skill development toward integrated learning experiences that connect language use with creativity, critical thinking, and reflective engagement. When embedded in coherent pedagogical structures, AI tools can support differentiation, learner autonomy, and sustained engagement, while preserving the central role of the teacher as a pedagogical guide whose feedback, instructional decisions, and professional judgment remain decisive for learning effectiveness (Hattie, 2009).

Second, teacher education and professional development should prioritize reflective competence in AI integration. Educators require not only technical familiarity with emerging tools, but also the pedagogical and ethical understanding necessary to evaluate their educational relevance and limitations. As argued by Godínez Martínez (2018), strengthening teachers' capacity for reflective practice is essential to ensure that technology remains responsive to learning goals.

Third, assessment practices in AI-enhanced learning environments should foreground formative and reflective approaches that capture linguistic development, collaboration, and critical engagement. Such assessment models, as recommended by Black & Wiliam (2018; 1998), support learner agency and emphasize learning processes alongside outcomes, allowing evaluation to function not as a summative mechanism but as an integral component of learning.

Recent OECD (OECD, 2026) analysis emphasizes that the educational value of generative artificial intelligence depends primarily on teachers' pedagogical mediation and its alignment with clear learning goals. Within this perspective, the lesson design presented in this paper illustrates how artificial intelligence can support language learning by fostering creativity, critical digital literacy, reflective engagement and social awareness. Artificial intelligence should therefore be understood as a pedagogical resource that expands educational possibilities while preserving the essential human dimensions of teaching.

References

Anastasiades, P., Kotsidis, K., Stratikopoulos, K. & Pananakakis, N. (2024). Human-Centered Artificial Intelligence in Education. The critical role of the educational community and the necessity of building a holistic pedagogical framework for the use of HCAI in education sector. *Open Education-The Journal for Open and Distance Education and Educational Technology, 20(1),* 29–51. https://doi.org/10.12681/jode.36612

Black, P. J., & Wiliam, D. (2018). Classroom assessment and pedagogy. *Assessment in Education: Principles, Policy & Practice, 25(1),* 1–25. https://doi.org/10.1080/0969594X.2018.1441807

Black, P.J., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7–74. https://doi.org/10.1080/0969595980050102

Brookfield, S. D. (1995). *Becoming a critically reflective teacher*. San Francisco, CA: Jossey-Bass.

Bruner, J. (1996). *The culture of education*. Cambridge, MA: Harvard University Press.

Council of Europe (2020) *Common European Framework of Reference for Languages: Learning, teaching, assessment-Companion volume*. Strasbourg: Council of Europe.

Dewey, J. (1938). *Experience and education*. N.Y: Macmillan.

European Commission. (2022). *Ethical guidelines on the use of artificial intelligence and data in teaching and learning for educators*. Luxembourg: Publications Office of the European Union.

Facer, K., & Selwyn, N. (2021). Digital Technology and the Futures of Education: Towards 'Non-Stupid' Optimism. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000377071

Fitzpatrick, S. (2026). Phase II: What needs to come next. Substack. https://fitzyhistory.substack.com/p/phase-ii-what-needs-to-come-next

Frøsig, T. B., & Romero, M. (2024). *Teacher agency in the age of generative AI: Towards a framework of hybrid intelligence for learning design.* arXiv. https://doi.org/10.48550/arXiv.2407.06655

Gee, J. P. (2015). *Social linguistics and literacies: Ideology in discourses* (5th ed.). https://doi.org/10.4324/9781315722511

Godínez Martínez, J. M. (2018). How effective is collaborative reflective practice in enabling cognitive transformation in English language teachers? *Reflective Practice, 19*(4), 427–446. https://doi.org/10.1080/14623943.2018.1479688

Godwin-Jones, R. (2024). Distributed agency in language learning and teaching through generative AI. Language Learning & Technology, 28(2), 5–30. https://doi.org/10.64152/10125/73570

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses related to achievement.* London: Routledge.

Hmelo-Silver, C. E. (2004). Problem-based learning: What and how do students learn? *Educational Psychology Review, 16*(3), 235–266. https://doi.org/10.1023/B:EDPR.0000034022.16470.f3,

Holmes, W., Bialik, M. and Fadel, C. (2019). *Artificial intelligence in education: Promises and implications for teaching and learning*. Boston, MA: Center for Curriculum Redesign.

Johnson, D. W., & Johnson, R. T. (2009). An educational psychology success story: Social interdependence theory and cooperative learning. *Educational Researcher, 38*(5), 365–379. https://doi.org/10.3102/0013189X09339057

Kern, R. (2015). Language, literacy, and technology. *Cambridge: Cambridge University Press.*

Knobel, M., & Lankshear, C. (2006). Digital Literacy and Digital Literacies: Policy, Pedagogy and Research Considerations for Education. *Nordic Journal of Digital Literacy, 1*(1), 12–24. https://doi.org/10.18261/ISSN1891-943X-2006-01-03

Kolb, D. A. (1984). *Experiential learning: Experience as the source of learning and development*. Englewood Cliffs, NJ: Prentice Hall.https://doi.org/10.4324/9780203970034

Kress, G. (2010). *Multimodality: A social semiotic approach to contemporary communication*. London: Routledge.

Lazcano-Quintana, I. (2024). Values-based education and the promotion of social participation: Evidence from educational leisure contexts in Spain. *Education Sciences, 14(4),* 430. https://doi.org/10.3390/educsci14040430

Luckin, R. (2018). *Machine learning and human intelligence: The future of education for the 21st century*. London: UCL Institute of Education Press.

Mishra, P., & Koehler, M. J. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. *Teachers College Record, 108*(6), 1017–1054. http://dx.doi.org/10.1111/j.1467-9620.2006.00684.x

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education, 31*(2), 199–218. https://doi.org/10.1080/03075070600572090

OECD (2026), *OECD Digital Education Outlook 2026: Exploring Effective Uses of Generative AI in Education*, OECD Publishing, Paris, https://doi.org/10.1787/062a7394-en,

OECD. (2023). *Artificial intelligence in education: Challenges and opportunities*. Paris: OECD Publishing. https://doi.org/10.1787/9c1d8c9f-en

Potkalitsky, N. (2025). *In praise of assistance*. Substack. https://nickpotkalitsky.substack.com

Puentedura, R.R. (2013). SAMR: Moving from enhancement to transformation. Available at: http://www.hippasus.com/rrpweblog/

Schön, D. A. (1983). *The reflective practitioner: How professionals think in action*. New York, NY: Basic Books. https://doi.org/10.4324/9781315237473

Schmager, S., Pappas, I. O., & Vassilakopoulou, P. (2025). Understanding human-centred AI: A review of its defining elements and a research agenda. Behaviour & Information Technology, https://doi.org/10.1080/0144929X.2024.2448719

Selwyn, N. (2019). *Should robots replace teachers? AI and the future of education*. Cambridge: Polity Press.

UNESCO. (2023). *Guidance for generative AI in education and research*. Paris: UNESCO Publishing. https://doi.org/10.54675/EWZM9535

UNESCO. (2021). *AI and education: Guidance for policy-makers*. Paris: UNESCO Publishing. https://doi.org/10.54675/PCSP7350

Vedder-Weiss, D., Roth, G. and Mishaeli, Y. (2025) 'Supporting teacher reflection and motivation through psychological needs satisfaction in collaborative reflection-based PD', *Journal of Experimental Education, 93*(2), pp. 320–339. https://doi.org/10.1080/00220973.2024.2309920

Voogt, J., Erstad, O., Dede, C. & Mishra, P. (2013). Challenges to learning and schooling in the digital networked world of the 21st century. *Journal of Computer Assisted Learning, 29*(5). https://doi.org/10.1111/jcal.12029

Vieriu, A. M., & Petrea, G. (2025). The Impact of Artificial Intelligence (AI) on Students' Academic Development. *Education Sciences, 15,* Article No. 343. https://doi.org/10.3390/educsci15030343

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

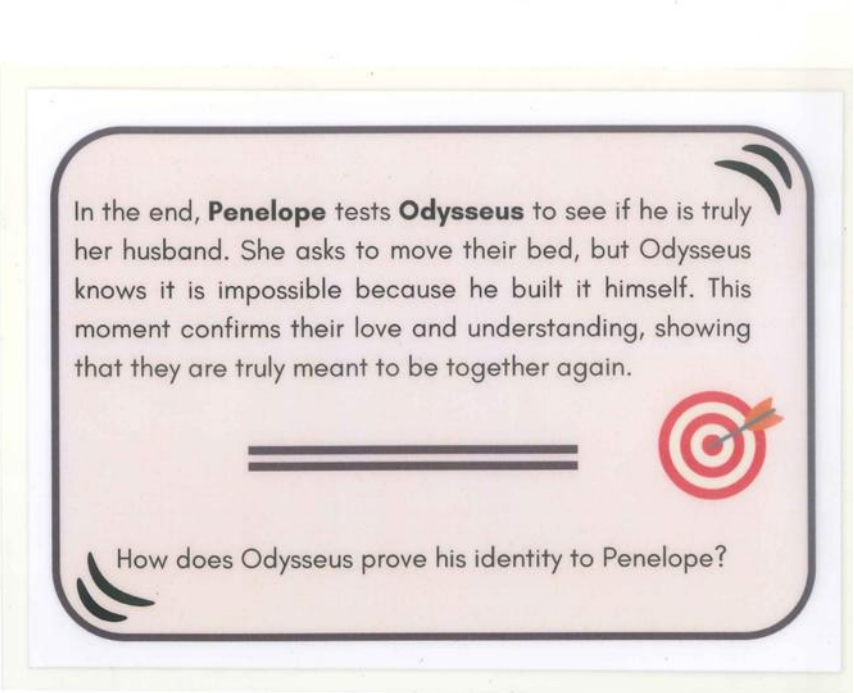
Warschauer, M., & Kern, R. (2000). *Network-based language teaching: Concepts and practice*. Cambridge: Cambridge University Press.

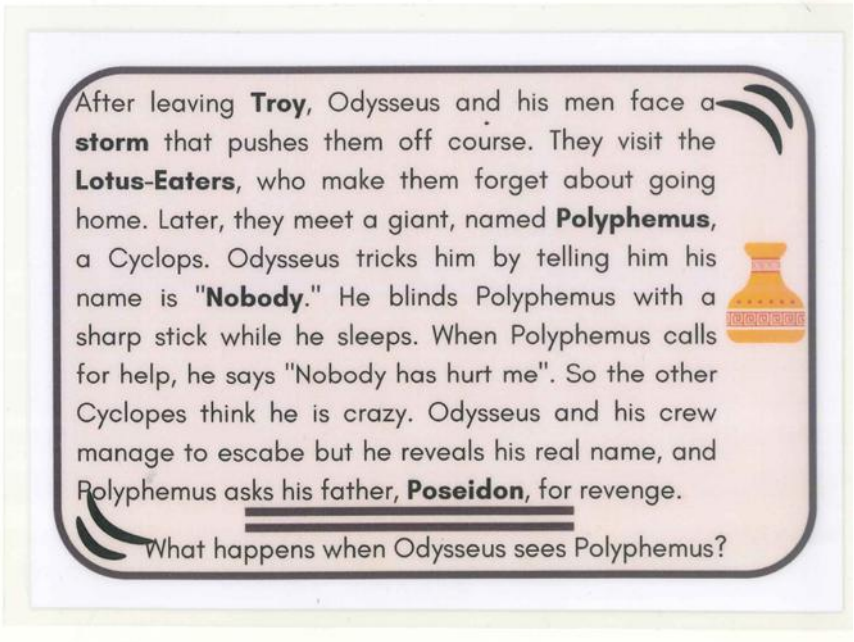
Williamson, B. (2016). Digital education governance: data visualization, predictive analytics, and "real-time" policy instruments. *Journal of Education Policy, 31*(2), 123–141. https://doi.org/10.1080/02680939.2015.1035758

Zhai, C. (2024). The effects of over-reliance on AI dialogue systems on students' cognitive abilities: A systematic review. *Smart Learning Environments, 11*(1). https://doi.org/10.1186/s40561-024-00316-7

Appendix A: Adapted Narrative Texts on Odysseus' Journey (Journey Timeline Activity)

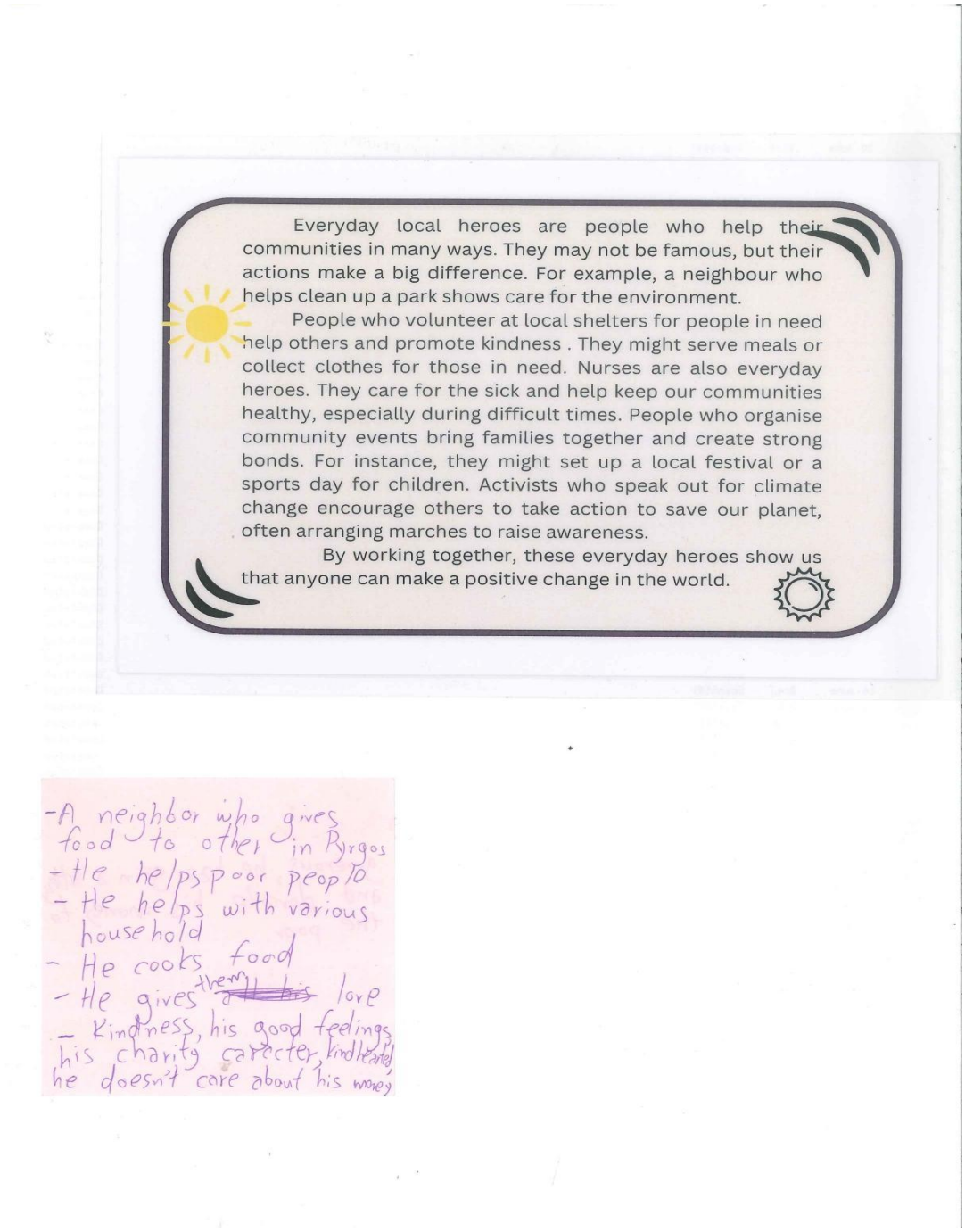
All student artifacts included in this appendix have been fully anonymised. Permission for their use was obtained in accordance with school and educational guidelines.





Appendix B: AI-Generated Informational Text and Sample Learner Notes on Community Heroes

All student artifacts presented in this appendix have been fully anonymised.



Everyday local heroes are people who help their communities in many ways. They may not be famous, but their actions make a big difference. For example, a neighbour who helps clean up a park shows care for the environment.

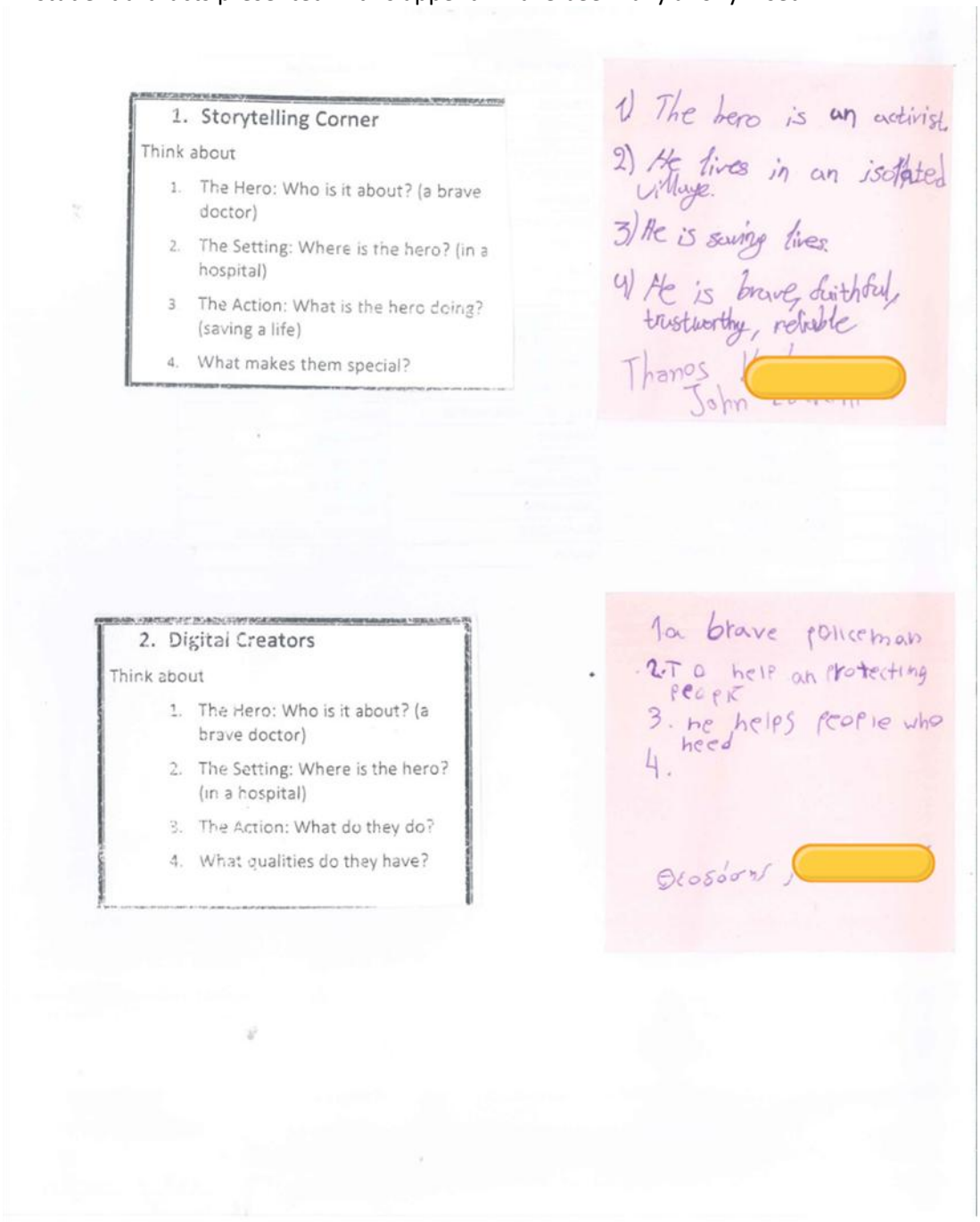
People who volunteer at local shelters for people in need help others and promote kindness . They might serve meals or collect clothes for those in need. Nurses are also everyday heroes. They care for the sick and help keep our communities healthy, especially during difficult times. People who organise community events bring families together and create strong bonds. For instance, they might set up a local festival or a sports day for children. Activists who speak out for climate change encourage others to take action to save our planet, often arranging marches to raise awareness.

By working together, these everyday heroes show us that anyone can make a positive change in the world.

-A neighbor who gives food to other in Pyrgos
-He helps poor people
- He helps with various household
- He cooks food
- He gives them ~~all his~~ love
- Kindness, his good feelings his charity carecter, kindhearted he doesn't care about his money

Appendix C: Examples of Anonymised Student Work from Learning Stations

All student artifacts presented in this appendix have been fully anonymised.



Appendix D: Examples of Anonymised Student Self-Assessment and Reflective Feedback

All student artifacts presented in this appendix have been fully anonymised.

Ρουμπρίκα Αυτοαξιολόγησης: Δημιουργία Βίντεο με Χρήση ΤΝ

Όνομα Μαθητή/τριας: [redacted]
Τάξη: [redacted]
Ημερομηνία: [redacted]

Σκέψου πώς εργάστηκες στην ομάδα σου και βάλε ✓στο κουτάκι που σε αντιπροσωπεύει:

Κριτήρια	Χρειάζομαι βοήθεια (1)	Προσπαθώ (2)	Τα πάω καλά (3)	Τα πάω πολύ καλά (4)
Χρήση Αγγλικής Γλώσσας Μίλησα στα Αγγλικά κατά τη συζήτηση για το βίντεο	Δυσκολεύτηκα να εκφραστώ στα Αγγλικά	Χρησιμοποίησα κάποιες αγγλικές εκφράσεις	Εκφράστηκα αρκετά στα Αγγλικά	Χρησιμοποίησα με άνεση τα Αγγλικά ✓
Χρήση Εργαλείου ΤΝ Χρησιμοποίησα το Lumen5 για τη δημιουργία του βίντεο	Δυσκολεύτηκα να χρησιμοποιήσω το Lumen5	Χρησιμοποίησα λίγο το Lumen5	Χρησιμοποίησα αρκετά το Lumen5	Χρησιμοποίησα με άνεση το Lumen5 ✓
Περιεχόμενο Το βίντεο παρουσιάζει έναν σύγχρονο ήρωα	Δυσκολεύτηκα να παρουσιάσω τον ήρωα	Παρουσίασα κάποια στοιχεία του ήρωα	Παρουσίασα αρκετά καλά τον ήρωα ✓	Παρουσίασα πολύ καλά τον ήρωα
Συμβολή στην Ομάδα Βοήθησα στην προσθήκη κειμένου/εικόνων στο βίντεο	Δυσκολεύτηκα να συνεισφέρω	Συνεισέφερα λίγο	Συνεισέφερα αρκετά ✓	Συνεισέφερα πολύ
Εστίαση Παρέμεινα συγκεντρωμένος/η στη δημιουργία του βίντεο	Δυσκολεύτηκα να συγκεντρωθώ	Ήμουν λίγο συγκεντρωμένος/η	Ήμουν αρκετά συγκεντρωμένος/η ✓	Ήμουν πολύ συγκεντρωμένος/η

Σχόλια μαθητή/τριας:

Μου άρεσε πολύ η δημιουργία του βίντεου επίσης είχαμε πλάκα να βρίσκουμε τα χαρακτηριστικά του ήρωα και εντυπωσιάστηκα που τόσο γρήγορα δημιουργήθηκε το βίντεο.

Sophia Kouzouli (sophiakouz@yahoo.com) is an Education Consultant for EFL teachers in the Directorates of Primary and Secondary Education of Ilia, Zakynthos, Kefalonia-Ithaca & Achaia. She holds a B.A. in English Language and Literature from the National and Kapodistrian University of Athens and a M.Ed. in Teaching English as a Foreign Language from the Hellenic Open University. She is a certified B-Level ICT trainer in Foreign Languages and has contributed to teacher professional development programmes organised by the Institute of Educational Policy. She has participated in the development of eTwinning MOOCs for teacher professional

learning. She is a member of the Greek National Support Organisation of the eTwinning action. Her professional and research interests include EFL pedagogy, artificial intelligence in education, educational technology, digital and critical literacy, values-based education, and teacher professional development.

Research Papers in Language Teaching and Learning
Vol. 16 No. 1, March 2026, 79-95
ISSN: 1792-1244
Available online at http://rpltl.eap.gr

GenAI-driven storytelling on senior secondary students' Generative AI literacy and writing skills: A mixed-methods research

Sezer Kizilates

The rapid development of Generative Artificial Intelligence (GenAI) is transforming education and requiring new forms of digital literacy. However, empirical research on the development of GenAI literacy in secondary education remains limited. This mixed-methods study investigated 76 students who were randomly assigned to control and experimental groups in the Hong Kong secondary school context. While the control group received traditional writing instruction, the experimental group engaged in a five-week GenAI-driven storytelling intervention. The study examined four dimensions of AI literacy, affective, behavioral, cognitive, and ethical, alongside key writing skills dimensions: word usage, narrative structure, creativity, and writing anxiety. Data were collected using validated questionnaires and a rubric-based writing assessment, administered before and after the intervention, as well as post-intervention interviews. Quantitative results showed that the experimental group reported reduced writing anxiety and achieved significant gains in AI literacy and writing performance. Thematic analysis confirmed these findings and highlighted greater motivation to use GenAI technologies, stronger confidence, and heightened ethical awareness. Overall, the study demonstrates the pedagogical value of integrating GenAI-driven storytelling into secondary education to foster AI literacy and enhance writing skills.

Keywords: AI literacy, digital storytelling, generative artificial intelligence (GenAI), narrative writing, creativity in writing, writing anxiety

1. Introduction

The integration of artificial intelligence (AI), particularly generative AI (GenAI) technologies, is significantly transforming the educational landscape. Since the launch of ChatGPT in late 2022, there has been an unprecedented surge of interest across the education sector in exploring how such technologies might reshape teaching and learning practices (Cacho, 2024). AI systems often emulate human cognitive functions—such as learning and problem-solving—by employing rule-based algorithms designed to process information, interpret inputs, and achieve specific objectives (Siemens et al., 2022). According to UNESCO (2023), GenAI refers to a subset of AI technologies that automatically generate content in response to prompts delivered through natural language-based conversational interfaces. More broadly, generative AI encompasses computational methods capable of producing novel and meaningful outputs—such as text, images, or audio—based on patterns learned from training data (Feuerriegel et al., 2023).

Traditional artificial intelligence (AI) and generative artificial intelligence (GenAI) exhibit substantial differences in both functionality and application. As outlined by Ng et al. (2025), six key dimensions distinguish GenAI from traditional AI: task execution, contextual understanding, response behavior, interactivity, data sources, and technical requirements. Traditional AI primarily relies on structured data derived from predefined datasets and historical records to generate conclusions (Ng et al., 2024). In contrast, GenAI is capable of generating human-like outputs across various modalities—including text, images, and videos. GenAI represents a rapidly expanding subfield of artificial intelligence that focuses on the development of models capable of generating new content—such as text, images, and music—with a level of originality that approximates human creativity. Beyond its generative capacity, GenAI empowers end users to create educational content independently (García-Peñalvo, 2023). Moreover, it enables the design of personalized learning experiences by adapting instructional difficulty and content in real time based on learner performance and behavioral data. GenAI can also support educators by generating instructional strategies and course materials informed by learners' progress and patterns of engagement (Huang, 2021).

Annapureddy *et al*. (2024) argue that the rapid advancement of Generative AI has rendered AI literacy increasingly essential. They highlight the need for specific competencies and knowledge to differentiate generative AI from other forms of AI and to understand how it produces creative outputs across domains such as writing, design, and scientific inquiry (Ray, 2023). Moreover, given that generative AI raises critical questions regarding authorship, ownership, and originality, ethical and legal literacy is equally important. As these technologies become more accessible to the general public, AI literacy must extend beyond technical experts to include broader societal engagement (Annapureddy et al., 2024). This view is echoed by Bozkurt (2024), who states that "its personality and essence will be shaped not only by the features it possesses but also by the capabilities and skills with which we use it." He further emphasizes that "the intentions and objectives of those who control it must also be taken into account." Within this framework, GenAI literacy assumes particular importance, as it fundamentally influences the evolving nature of human–machine interaction.

Despite the increasing incorporation of GenAI tools into educational environments, current literature has predominantly concentrated on higher education, focusing on theoretical models, tool functionalities, and general ethical considerations (Annapureddy et al., 2024; Bozkurt, 2024). In contrast, there is a notable scarcity of empirical, classroom-based investigations that explore how GenAI literacy can be cultivated specifically among secondary school students. This study aims to examine the impact of a GenAI-driven storytelling workshop on the development of generative AI literacy and writing proficiency among secondary students. Specifically, it investigates whether the intervention can enhance students' GenAI literacy across four key domains: affective, behavioral, cognitive, and ethical. In addition, the study explores how students perceive these changes and how GenAI-assisted storytelling may influence their writing performance in terms of word usage, narrative structure, creativity, and writing anxiety.

The study aims to answer three research questions:

1. To what extent does GenAI-driven storytelling support the development of students' generative AI literacy across affective, behavioral, cognitive, and ethical dimensions?
2. How does GenAI-driven storytelling impact students' GenAI literacy in terms of affective, behavioral, cognitive, and ethical dimensions?
3. To what extent does GenAI-driven storytelling influence students' writing skills in terms of word usage, narrative structure, creativity, and writing anxiety?
4. How do students perceive the impact of GenAI-driven storytelling on their writing skills, particularly in terms of word usage, narrative structure, creativity, and writing anxiety?

2. Literature Review

2.1 Defining AI Literacy and Its Evolution

The term "AI literacy" refers to the understanding and ability to effectively engage with artificial intelligence technologies. It encompasses a broad set of competencies necessary for navigating an AI-driven world. Core components include conceptual knowledge of AI, operational skills in using AI tools, and the critical capacity to evaluate the implications of AI applications across diverse contexts (Ng et al., 2021). A growing body of literature suggests that AI literacy is an evolving and multifaceted construct. It extends beyond mere technical proficiency to include ethical considerations and social impacts, implying that individuals must be prepared to understand and respond to the broader consequences of AI in both community and everyday life (Biagini, 2025).

2.2 From AI Literacy to GenAI Literacy: Key GenAI Literacy Frameworks

Following the introduction of various AI literacy guidelines, the proliferation of generative AI tools has radically transformed daily life. GenAI presents distinctive and more complex characteristics compared to traditional AI algorithms, which are not fully addressed in current AI literacy frameworks (Zhang & Magerko, 2025). The emergence of GenAI tools such as ChatGPT necessitates a specific set of competencies that extend beyond general AI knowledge, and it is fundamentally reshaping the

educational landscape, presenting both significant opportunities and complex challenges (Annapureddy et al., 2024; Bozkurt, 2024). Zhang and Magerko (2025) offer a flexible, practice-oriented framework comprising twelve principles for fostering generative AI literacy, grounded in a comprehensive analysis of existing scholarship.

Likewise, Bozkurt's (2024) 3wAI framework introduces a metaphorically inspired, conceptually rich approach to building GenAI literacy. Rather than prescribing rigid competencies, the model is designed to be adaptive, contextually responsive, and scalable across diverse user groups. In parallel, Annapureddy, Fornaroli & Gatica-Perez (2025) have proposed a structured, competency-based framework that accounts for the complexity and dynamic evolution of GenAI technologies.

2.3 GenAI Applications in Education

As a sophisticated technology with significant immersive potential, generative AI (GenAI) presents both opportunities and challenges within educational contexts (Tlili et al., 2023). On the positive side, GenAI enhances student engagement through personalized feedback and interactive learning experiences. It adapts learning pathways in real time based on student performance, providing immediate support that fosters deeper understanding (Anderson et al., 2025). Lo (2023) demonstrates that ChatGPT can function as both a virtual tutor and instructional assistant, fulfilling a wide array of educational needs. Expanding on this perspective, tools such as ChatGPT and Claude have been identified as valuable assets in supporting personalized learning experiences. These technologies equip students and their families with strategies to navigate contemporary educational demands more effectively (Narciso, 2024).

Nevertheless, the limitations of GenAI must be acknowledged. These include its inability to fully interpret real-world contexts and its potential to produce inaccurate information (Chiu, 2023). Concerns have also been raised regarding the potential erosion of learners' critical thinking skills and the risk of over-reliance on AI systems. Additionally, educators have highlighted privacy-related issues, particularly the misuse of data and unauthorized access that may compromise the security of student information. The credibility of academic work has also been questioned due to GenAI's potential to facilitate plagiarism and diminish academic rigor (Pikhart & Al-Obaydi, 2025). Ethical implications are another recurring theme. In addition to calls for comprehensive ethical guidelines to govern AI usage, scholars have expressed concerns about algorithmic bias and the environmental impact of large-scale AI deployment (Pitts et al., 2025).

2.4 Conceptual Rationale for Digital Story Writing (DSW)

The adoption of GenAI-driven digital story writing (DSW) in this study was grounded in established constructionist and inquiry-based pedagogical traditions rather than being motivated solely by technological novelty. DSW was selected because constructionist approaches emphasize learning through making meaningful digital artefacts using technology (Hughes et al., 2017; Laurillard et al., 2011). In DSW, students actively construct stories by integrating text, images, and other media, which allows them to reinterpret newly learned AI concepts rather than simply receiving information (Woo et al., 2013). This

process aligns with inquiry-based learning, where reading and writing are used to explore information, reorganize knowledge, and deepen conceptual understanding through reflection and creation (McGinley & Tierney, 1988; Chu et al., 2021). In the context of AI learning, story writing requires students to first observe and understand AI concepts, then organize, classify, and apply them within a meaningful narrative, supporting deeper cognitive processing (Wong et al., 2020).

DSW was also chosen because it supports multimodal learning by allowing students to express ideas through multiple modes, including text, images, audio, and video. This multimodal approach helps make abstract concepts more accessible and supports learners with varying language proficiency levels (Boase, 2008; Smyrnaiou et al., 2020; Eisenlauer & Karatza, 2020). Prior research has shown that DSW provides a strong foundation for multimodal literacies by shifting learning from purely verbal modes to multimedia forms of expression (Ng et al., 2024). By engaging with different semiotic resources during story creation, students are able to connect prior knowledge with new concepts and construct meaning through the production of a finalized digital story (Boase, 2008).

3. Methodology

3.1 Research Context

This study was conducted at a Band 3 secondary school in Hong Kong and involved a total of 76 students from Form 4 and Form 5. The participants were divided into two groups: an experimental group (n = 38), which participated in a five-week, ten-hour GenAI-driven storytelling workshop, and a control group (n = 38), which continued with conventional English lessons without exposure to any AI tools.

This study was conducted in a government-funded secondary school in a public housing district serving socioeconomically disadvantaged students. The school had limited digital resources, while students exhibited low levels of digital and GenAI literacy. Many students lacked access to personal computers and had limited keyboarding skills, constraining their engagement in extended digital writing tasks. Institutionally, there was no school-wide AI policy, generative AI was not embedded in the curriculum, and teachers reported low confidence and limited pedagogical experience with AI tools.

3.2 GenAI Storytelling Course Design and Theoretical Framework

The intervention spanned five consecutive weeks, with two 60-minute sessions per week, conducted in the school's computer lab. Each session followed a structured instructional sequence grounded in the 5-step GenAI storytelling model, aligned with the study’s research questions, and supported by relevant pedagogical and AI literacy frameworks. The structure of the GenAI storytelling model is presented in Table 1.

The design of the intervention was guided primarily by Ng et al.'s (2024) AI Literacy Framework, which conceptualizes AI literacy across four key dimensions: affective, behavioral, cognitive, and ethical. Each

workshop session was intentionally designed to target one or more of these dimensions through scaffolded GenAI storytelling tasks. See Table 1 for details.

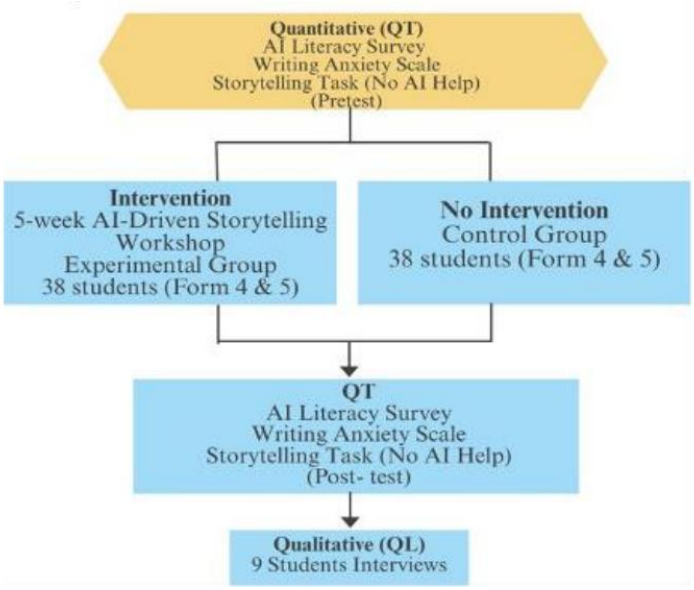


Figure 1: The overall research design.

The intervention was implemented as a pedagogically scaffolded process. Initial sessions introduced basic AI concepts and appropriate ways of interacting with generative tools. During the storytelling tasks, students were provided with instructor-designed prompt banks, along with guided prompt examples, individual and small-group support, and targeted assistance for lower-performing learners. The prompt banks supported idea generation, offered opportunities for brainstorming, and encouraged creative exploration during story production.

3.3 Research Design, Data Collection, and Sampling Method

The study employed a pretest-posttest quasi-experimental design (see Figure 1) and adopted a mixed-methods research approach. Quantitative data were collected using standardized questionnaires and rubric-based assessments of student writing, while qualitative data were obtained through semi-structured interviews. These instruments were administered to both the experimental and control groups immediately before and after the five-week intervention. Participants were assigned to either the experimental or control group through a process of simple random sampling.

Week	Focus	Key Activities	AI Literacy Outcomes	Writing Outcomes
1	What is AI? / AI vs GenAI	Introduction to AI concepts, classroom	Affective, Ethical	Writing Anxiety (building trust in AI)

		discussion, comparison of AI- vs human-generated stories		
2	Text Generation / Prompt Crafting	5W1H brainstorming, prompt crafting, and outline building using Gen AI tools	Cognitive, Behavioral	Narrative Structure, Creativity, Word Usage
3	Digital Arts Creation (Visuals)	Create and revise visuals using Gen AI tools, and integrate visuals into storylines	Behavioral	Narrative Structure, Creativity
4	AI Music Composition	Generate and select AI music, align sound with narrative mood, and conduct ethical reflection	Affective, Ethical, Cognitive	Narrative Structure
5	Final Story Creation & Presentation	Story editing, peer review, highlight AI parts, record narration, and present	Ethical, Behavioral	Word Usage, Writing Anxiety

Table 1: 5-step GenAI storytelling model

3.3.1 Quantitative Data Instruments

Three validated instruments were used to collect data:

a) AI Literacy Questionnaire (Ng et al., 2024): This survey assessed students' AI literacy across four dimensions— affective, behavioral, cognitive, and ethical.

b) Writing Anxiety Scale (Cheng, 2004): To examine the emotional aspects of writing, this scale measured students' levels of anxiety related to writing tasks.

c) Creative Writing Assessment Rubric (Carey et al., 2022): A performance-based assessment was employed to evaluate students' writing samples on the dimensions of word usage,
narrative structure, and creativity.

3.3.2 Qualitative Data

To gain deeper insight into students' experiences, semi-structured interviews were conducted with selected students from the experimental group after the intervention. Interview questions explored their attitudes toward GenAI tools, perceived improvements in writing, emotional responses, and ethical considerations. Interviews lasted approximately 20–30 minutes and were audio-recorded for transcription and analysis. The interviews helped contextualize quantitative findings and shed light on how students engaged with GenAI in a real learning environment.

3.6 Data Analysis

To evaluate the effects of the GenAI-driven storytelling intervention and answer the research questions, both quantitative and qualitative data were analyzed using appropriate statistical and interpretive methods. Quantitative data obtained from the AI Literacy Questionnaire, Writing Anxiety Scale, and Creative Writing Assessment were analyzed using SPSS. Descriptive statistics, including means and standard deviations, were first calculated to summarize pre- and post-test scores for both experimental and control groups. To assess the equivalency of groups before the intervention, independent samples t-tests were conducted. Within-group changes from pre- to post-test were examined using paired samples t-tests. The pre-test comparisons indicated no statistically significant differences between the groups. ANCOVA was employed when comparing post-test scores to control for potential baseline variability and to increase the precision of the group comparisons..In addition, Cohen's d was calculated to measure the effect size of the intervention, providing insight into the significance of observed changes.

Thematic analysis, utilising the six-step framework developed by Braun and Clarke (2006), was used for qualitative analysis. Transcripts of interviews with students in the selected experimental group were openly coded to find recurring patterns. The codes were examined, categorized, and precisely defined into more general themes related to writing outputs, including word usage, narrative structure, creativity, and writing anxiety, as well as the affective, behavioral, cognitive, and ethical aspects of AI literacy.

4. Results and Findings

4.1 Background and Pre-Test Equivalence

A total of 76 students were divided equally into experimental and control groups. The experimental group included 18 Form 4 students (47.4%) and 20 Form 5 students (52.6%), ensuring a balanced distribution across grade levels. Pre-test comparisons confirmed that the two groups were statistically equivalent across all GenAI literacy domains, including affective, behavioral, cognitive, and ethical dimensions, as well as in writing skills (word usage, narrative structure, creativity, and writing anxiety). This baseline equivalence supported the validity of using post-test differences to assess intervention effects.

RQ1: To What Extent Does GenAI-driven Storytelling Impact Students' GenAI Literacy?

Post-test analyses indicated higher GenAI literacy scores across affective, behavioral, cognitive, and ethical dimensions in the experimental group following the intervention. Mann–Whitney U tests revealed no statistically significant differences between Form 4 and Form 5 students at either the pre-test or post-test stages across any GenAI literacy domains ($p > .05$). Both grade levels showed increased GenAI literacy scores after participation in the GenAI-driven storytelling workshop, as illustrated in Figure 2. In contrast, the control group did not demonstrate statistically significant changes in any GenAI literacy dimension between the pre-test and post-test stages (all $p > .05$).

Findings: Statistical results showed significant gains across all four GenAI literacy dimensions in the experimental group ($p < .001$). No statistically significant differences were observed between Form 4 and Form 5 students at either testing point ($p > .05$), and no corresponding gains were identified in the control group.

RQ2: How Does GenAI-driven Storytelling Impact Students' GenAI Literacy?

The thematic analysis revealed a set of sub-themes across the dimensions, capturing different aspects of students' engagement with GenAI-driven storytelling.

1. Affective Domain
    - **Emotional shift.** Students described moving from initial hesitation toward more positive attitudes. "At the start, I was a bit skeptical… but I realized you can still have creativity" (S2).
    - **Increased confidence.** Students associated AI-supported storytelling with a greater sense of control, emotional reassurance, and confidence during writing tasks. "I felt successful because I made a story… I felt proud of myself" (S1).
    - **Enhanced motivation.** Students expressed curiosity and enjoyment when experimenting with different AI features. "I didn't know I can use my voice to give AI the rhythm while creating music… AI helped me make it" (S3).
2. Cognitive Domain
    - **Idea development.** Students demonstrated cognitive awareness of AI as an idea-generation catalyst rather than an idea owner. "I had a story about a notebook... AI gave me different perspectives and how I could elaborate on different parts." (S2)
    - **Integration of AI and personal input.** Students described actively merging their own ideas with AI-generated suggestions to enrich characters, settings, and plot elements. "I thought of alien invasion… but AI helped me elaborate like aliens from another universe, with new characters and background." (S1)
    - **Language awareness.** Increased awareness of vocabulary and sentence-level features was evident. "I wrote down 'reputed' and 'obtuse.' I didn't know these before, but AI helped me learn such words." (S1)
3. Behavioral Domain
    - **Strategic integration of AI.** Students described using GenAI purposefully across different stages of writing. "First I will give the opening… then the problem… then the ending" (S3).
    - **Extension of AI use beyond the classroom.** Students reported applying GenAI tools in other academic subjects and informal contexts. "I used it to help with astronomy class… and also in economics class" (S5).

- **Autonomous prompting.** Increased procedural awareness was evident as students described refining prompts and selectively improving AI outputs. "I give it an idea, then refine what it gives back" (S2).
- **Collaborative learning.** Some students described using GenAI in collaborative contexts. "We shared our main ideas, and AI helped us create the story" (S1).

4. Ethical Domain

- **Using AI as support.** Students emphasized that effective AI use required intentional personal input, positioning GenAI as a tool that supports rather than replaces their own thinking. "I did not let it make a story. I definitely made my own story, and the ideas it gave me, I changed." (S2)
- **Authorship awareness and transparency.** Students emphasized the importance of originality and honesty regarding AI involvement. "When you use AI, you should be honest about it… tell what part was written by AI, what part is written by you" (S9).
- **Evolving ethical understanding.** Students described a progression from guilt toward clearer ethical judgment. "At first, I thought using AI was kind of like cheating… but now I know if I give it my ideas, it helps me learn and improve" (S2).

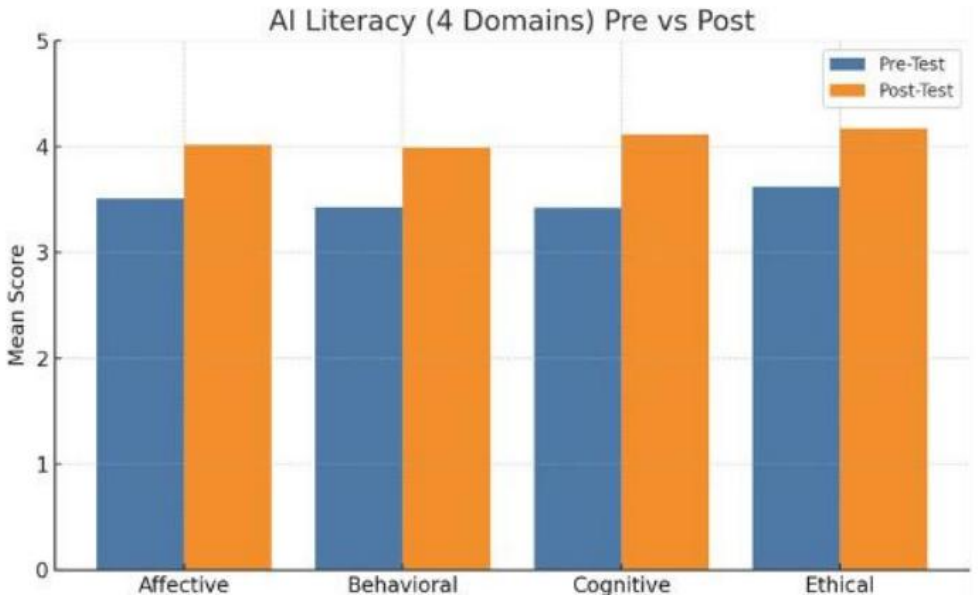


Figure 2: Experimental Group Comparison of Pre- and Post-Test Scores Across Four GenAI Literacy Domains across grade levels.

RQ3: To What Extent Does GenAI-driven Storytelling Influence Students' Writing Skills?

Post-test results indicated higher writing scores for students in the experimental group following participation in the GenAI-driven storytelling workshop. The mean overall writing score increased from 3.34 to 8.71 out of 15. Improvements were observed across multiple writing dimensions, including word usage, narrative structure, and creativity, as presented in Figure 3. In contrast, the control group did not

show statistically significant changes in overall writing performance or writing anxiety across any measured dimension between the pre-test and post-test stages.

Findings: Results showed significant gains across all assessed writing skill dimensions in the experimental group following the intervention, as shown in Figure 3. In addition, statistically significant reductions were observed across all three subdimensions of writing anxiety ($p < .001$), as illustrated in Figure 4. No corresponding improvements were observed in the control group.

RQ4: How Do Students Perceive the Impact of GenAI-driven Storytelling on Their Writing Skills?

The thematic analysis revealed sub-themes capturing students' perceptions of how GenAI-driven storytelling influenced their writing skills across word usage, creativity, narrative structure, and writing anxiety.

1. Word Usage
    - **Vocabulary expansion.** Students reported acquiring new and unfamiliar words through AI-generated suggestions. "It gave me good words that I didn't know… and I used them." (S3).
    - **Lexical appropriacy awareness.** Students demonstrated increased awareness of word choice. "The AI gave many versions, then I chose one that was more suitable" (S7).
    - **Motivation to experiment with words.** Students described greater willingness to try new and varied vocabulary, often framing language use as playful exploration supported by AI. "I tried to write more interesting words because AI gave me some cool options" (S8).
2. Creativity
    - **Idea expansion.** Students reported that AI-generated input supported the development of more imaginative and detailed storylines by introducing unexpected or novel ideas. "AI gave alien ideas, I used them with my own story" (S6).
    - **Blending human and AI imagination.** Students described combining personal ideas with AI-generated suggestions. "The jungle and characters were mine… but I asked AI to help with the setting" (S5).
    - **Playful and experimental attitude.** Students framed AI-supported storytelling as an enjoyable and exploratory process, often treating it as a creative game. "It felt like a game… using AI to create something strange" (S1).

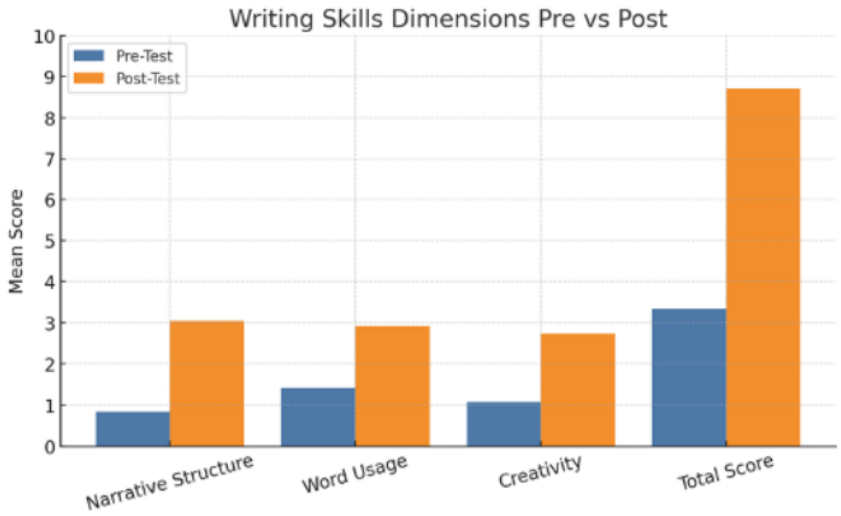


Figure 3: Experimental Group Comparison of Pre- and Post-Test Scores Across Three Storytelling Skills Domains

3. Narrative Structure

- **Narrative structure awareness and scaffolding.** Students demonstrated improved understanding of core narrative elements and used GenAI to organise story components more effectively. "I had only some parts… AI helped me organize them better" (S7).
- **Increased coherence and flow.** Students reported improved logical sequencing of ideas, smoother transitions, and greater overall coherence. "The AI helped make it more connected… it flowed better than my first draft" (S6).

4. Writing Anxiety

- **Reduction of writing-related stress.** Students reported feeling less stressed and more supported, describing AI as a source of reassurance that made writing more approachable. "I feel less stressed because AI gives me a start" (S8).
- **Increased confidence to take linguistic risks.** Students described greater willingness to experiment with language and ideas. "Before I think my writing is weak… now I feel more brave" (S7).
- **Transformed anxiety.** Some students continued to express concerns about overuse, dependence, or authorship, indicating that anxiety was not entirely eliminated but reshaped. "Sometimes I still feel worried… like, is this really my idea?" (S3).

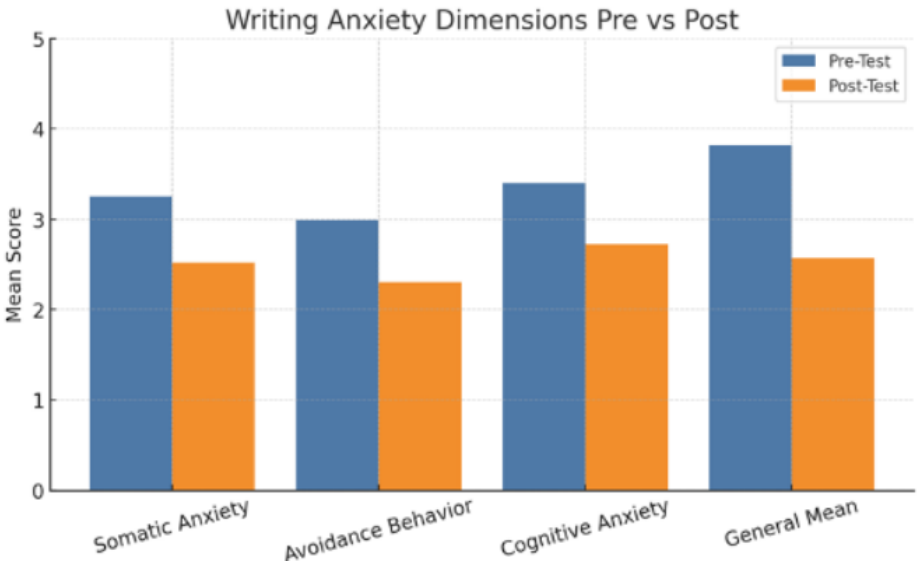


Figure 4: Pre- and Post-Test Scores of Writing Anxiety Subdimensions in the Experimental Group

5. Discussion

The findings indicate that the GenAI-driven storytelling workshop enhanced students' writing skills and AI literacy in a Band 3 secondary school context characterised by limited digital infrastructure and low initial AI literacy. Beyond overall gains, the qualitative findings help explain how these developments unfolded across GenAI literacy dimensions. Affective changes suggest that GenAI-driven storytelling reduced initial resistance to AI use and supported emotional engagement, creating conditions in which students felt more confident and motivated to participate in AI-supported writing. Cognitive engagement indicates a shift from viewing AI as an answer-providing tool toward recognising it as a generative resource that requires human judgment, particularly in idea development, language use, and narrative construction. Behavioral patterns further suggest increasing learner agency, as students moved toward more strategic, autonomous, and transferable uses of GenAI across writing stages and learning contexts. Ethical reflections point to an emerging understanding of authorship and responsibility, indicating that students did not adopt GenAI uncritically but actively negotiated its role in relation to originality and personal contribution. Collectively, these patterns suggest that GenAI literacy developed as a reflective and interconnected process rather than as isolated affective, cognitive, behavioral, or ethical skills.

Parallel to these literacy-related developments, students' perceptions of their writing skills suggest that GenAI-driven storytelling reshaped the writing experience in focused yet interconnected ways. Students described engaging with language in a more exploratory and intentional manner, experimenting with vocabulary choices while remaining attentive to meaning and appropriacy. The writing process also became more generative, as GenAI support enabled students to extend, adapt, and elaborate ideas without displacing their original creative intentions. At the level of text organisation, GenAI was perceived as a scaffold that supported clearer structuring of drafts, helping students manage gaps, strengthen sequencing, and improve overall coherence. Although writing-related stress was reduced and confidence

increased, students continued to reflect on issues of authorship and reliance, indicating that emotional reassurance coexisted with growing critical awareness rather than unexamined dependence.

The results both align with and further existing research. For instance, Bozkurt and Sharma (2024) contend that ethical AI integration in education requires explainability, transparency, and trust. The increased ethical awareness observed in this study—particularly students' understanding of authorship, attribution, and the responsible use of AI—echoes Bozkurt's emphasis on trustworthy and interpretable AI practices. The cognitive and behavioral improvements observed are also in line with the findings of Annapureddy et al. (2024). A notable reduction in writing anxiety further highlights GenAI's role not only as a cognitive aid but also as an affective support mechanism, echoing findings by Hawanti and Zubaydulloevna (2023) regarding AI's potential to foster learner confidence and engagement.

Importantly, these findings reinforce the notion that GenAI literacy constitutes a recursive and reflective process rather than a linear progression of discrete skills, aligning with the hybrid literacy paradigm articulated by Bozkurt (2024). The emergence of ethical awareness, for example, signals a developing conceptualization of AI not merely as a tool, but as a collaborative agent whose role must be continually examined and negotiated. This conceptual deepening suggests that pedagogically guided GenAI use can support students' evolving understanding of agency, responsibility, and authorship in AI-mediated writing contexts. Such insights further substantiate the inclusion of affective and cognitive dimensions in future GenAI literacy models, thereby validating the framework advanced by Ng et al. (2024).

As a practical and empirical contribution, the findings show that GenAI can be integrated into everyday classroom practice by embedding it within structured writing tasks rather than using it as a stand-alone tool for teachers. Guiding students through idea generation, drafting, revision, and reflection allows GenAI to support writing development and AI literacy in a focused and manageable way. Importantly, this approach demonstrates that meaningful GenAI integration does not require advanced technical infrastructure, but rather clear pedagogical goals, scaffolded activities, and opportunities for reflection, making it particularly relevant for resource-constrained secondary school contexts.

6. Conclusion

This study examined the impact of a GenAI-driven storytelling workshop on secondary school students' writing skills and generative AI literacy in a Hong Kong secondary school context. Using a mixed-methods approach, the study explored changes in students' writing performance, including word usage, narrative structure, creativity, and writing anxiety, and engagement with GenAI across affective, behavioral, cognitive, and ethical dimensions. The findings indicate that GenAI-driven storytelling can support students' writing development while also strengthening their understanding and responsible use of generative AI tools. Students showed clearer awareness of narrative structure, richer word use, increased creativity, and reduced writing anxiety. At the same time, they developed greater confidence, motivation, and ethical awareness when interacting with GenAI technologies.

Overall, this study points to the potential for developing key AI literacy competencies within resource-constrained educational contexts when supported by structured pedagogical interventions. Students can engage productively with GenAI when learning activities emphasize creativity, reflection, and guided use rather than technical complexity. The results should be interpreted in light of several limitations, including the short intervention period and the use of a single school context, which may limit generalizability. These findings highlight the value of integrating GenAI-supported story writing into English language education and point to the need for further classroom-based research in diverse secondary school settings.

References

Anderson, J. E., Nguyen, C. A., & Moreira, G. (2025). Generative AI-driven personalization of the Community of Inquiry model: enhancing individualized learning experiences in digital classrooms. *International Journal of Information and Learning Technology*, 42(3), 296–310. https://doi.org/10.1108/ijilt-10-2024-0240

Annapureddy, R., Fornaroli, A., & Gatica-Perez, D. (2024). Generative AI Literacy: Twelve defining competencies. *Digital Government Research and Practice*, 6(1), 1–21. https://doi.org/10.1145/3685680

Biagini, G. (2025). Towards an AI-Literate Future: A systematic literature review exploring education, ethics, and applications. *International Journal of Artificial Intelligence in Education*, 35(4), 2616–2666. https://doi.org/10.1007/s40593-025-00466-w

Boase, J. (2008). Personal networks and the personal communication system: Using multiple media to connect. *Information, Communication & Society*, 11(4), 490–508. https://doi.org/10.1080/13691180801999001

Bozkurt, A. (2024). Why Generative AI Literacy, Why Now, and Why it Matters in the Educational Landscape? Kings, Queens, and GenAI Dragons. *Open Praxis*, 16(3), 283–290. https://doi.org/10.55982/openpraxis.16.3.739

Bozkurt, A., Sharma, R. C. (2024). Trust, credibility, and transparency in human-AI interaction: Why we need explainable and trustworthy AI and why we need it now?. *Asian Journal of Distance Education, 19(*2). Retrieved from https://www.asianjde.com/ojs/index.php/AsianJDE/article/view/819

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Cacho, R. (2024). Integrating Generative AI in university Teaching and Learning: a model for balanced guidelines. *Online Learning*, 28(3). https://doi.org/10.24059/olj.v28i3.4508

Carey, M. D., Davidow, S., & Williams, P. (2022). Re-imagining narrative writing and assessment: A post-NAPLAN craft-based rubric for creative writing. *Australian Journal of Language and Literacy*, 45(1), 33–48. https://doi.org/10.1007/s44020-022-00004-4

Cheng, Y. (2004). A measure of second language writing anxiety: Scale development and preliminary validation. *Journal of Second Language Writing*, 13(4), 313–335. https://doi.org/10.1016/j.jslw.2004.07.001

Chiu, T. K. F. (2023). The impact of Generative AI (GenAI) on practices, policies and research direction in education: a case of ChatGPT and Midjourney. *Interactive Learning Environments*, 32(10), 6187–6203. https://doi.org/10.1080/10494820.2023.2253861

Chu, S. K. W., Reynolds, R. B., Tavares, N. J., Notari, M., & Lee, C. W. Y. (2021). *21st-century skills development through inquiry-based learning from theory to practice*. Springer International Publishing. https://doi.org/10.1007/978-981-10-2481-8

Eisenlauer, V., & Karatza, S. (2020). Multimodal literacies: Media affordances, semiotic resources and discourse communities. *Journal of Visual Literacy*, 39(3–4), 125–131. https://doi.org/10.1080/1051144x.2020.1826224

Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2023). Generative AI. *Business & Information Systems Engineering*, 66(1), 111–126. https://doi.org/10.1007/s12599-023-00834-7

García-Peñalvo, F. J. (2023). The perception of Artificial Intelligence in educational contexts after the launch of ChatGPT: Disruption or Panic? *Education in the Knowledge Society*, 24, Article e31279. https://doi.org/10.14201/eks.31279

Hawanti, S., & Zubaydulloevna, M. (2023). AI chatbot-based learning: Alleviating students' anxiety in English writing classroom. *Bulletin of Social Informatics Theory and Application*, 7(2), 182–192. https://doi.org/10.31763/businta.v7i2.659

Huang, X. (2021). Aims for cultivating students' key competencies based on artificial intelligence education in China. *Educational Information Technology*, 26, 5127–5147. https://doi.org/10.1007/s10639-021-10530-2

Hughes, J., Gadanidis, G., & Yiu, C. (2017). Digital making in elementary mathematics education. *Digital Experiences in Mathematics Education*, 3(2), 139–153. https://doi.org/10.1007/s40751-016-0020-x

Laurillard, D., Charlton, P., Craft, B., Dimakopoulos, D., Ljubojevic, D., Magoulas, G., Masterman, E.Pujadas, R., Whitley, E., & Whittlestone, K. (2011). A constructionist learning environment for teachers to model learning designs. *Journal of Computer Assisted Learning*, 29(1), 15–30. https://doi.org/10.1111/j.1365-2729.2011.00458.x

Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences, 13*(4), 410. https://doi.org/10.3390/educsci13040410

McGinley, W., & Tierney, R. J. (1988). Reading and writing as ways of knowing and learning. *Center for the Study of Reading Technical Report*. No. 423. https://eric.ed.gov/?id=ED294136

Narciso, P. (2024). *Generative AI in education: A guide for parents and teachers*. Springer Nature. https://doi.org/10.1007/979-8-8688-0844-9

Ng, D. T. K., Chan, E. K. C., & Lo, C. K. (2025). Opportunities, challenges, and school strategies for integrating generative AI in education. *Computers & Education: Artificial Intelligence*, 100373. https://doi.org/10.1016/j.caeai.2025.100373

Ng, D. T. K., Leung, J. K. L., Chu, K. W. S., & Qiao, M. S. (2021). AI literacy: Definition, teaching, evaluation, and ethical issues. *Proceedings of the Association for Information Science and Technology*, 58(1), 504–509. https://doi.org/10.1002/pra2.487

Ng, D. T. K., Wu, W., Leung, J. K. L., Chiu, T. K. F., & Chu, S. K. W. (2024). Design and validation of the AI literacy questionnaire: The affective, behavioural, cognitive and ethical approach. *British Journal of Educational Technology*, 55, 1082–1104. https://doi.org/10.1111/bjet.13411 Pikhart, M., & Al-Obaydi, L. H. (2025). Reporting the potential risk of using AI in higher education: Subjective perspectives of educators. *Computers in Human Behavior Reports*, 18, 100693. https://doi.org/10.1016/j.chbr.2025.100693

Pitts, G., Marcus, V., & Motamedi, S. (2025). *Student perspectives on the benefits and risks of AI in education*. arXiv. https://doi.org/10.48550/arxiv.2505.02198

Shen, X., & Tao, Y. (2025). Metacognitive strategies, AI-based writing self-efficacy, and writing anxiety in AI-assisted writing contexts: A structural equation modeling analysis. *International Journal of TESOL Studies, 7(*1), 70–87. https://doi.org/10.58304/ijts.20250105

Siemens, G., Marmolejo-Ramos, F., Gabriel, F., Medeiros, K., Marrone, R., Joksimovic, S., & de Laat, M. (2022). Human and artificial cognition. *Computers & Education: Artificial Intelligence*, 3, 100107. https://doi.org/10.1016/j.caeai.2022.100107

Smyrnaiou, Z., Georgakopoulou, E., & Sotiriou, S. (2020). Promoting a mixed-design model of scientific creativity through digital storytelling—the CCQ model for creativity. *International Journal of STEM Education*, 7(1). https://doi.org/10.1186/s40594-020-00223-6

Tlili, A., Shehata, B., Adarkwah, M. A., Bozkurt, A., Hickey, D. T., Huang, R., & Agyemang, B. (2023). What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart Learning Environments*, 10(1). https://doi.org/10.1186/s40561-023-00237-x

UNESCO (2023). *Guidance for Generative AI in Education and Research*. https://www.unesco.org/en/articles/guidance-generative-ai-education-and-research

Wong, G. K. W., Ma, X., Dillenbourg, P., & Huan, J. (2020). Broadening artificial intelligence education in K-12. *ACM Inroads*, 11(1), 20–29. https://doi.org/10.1145/3381884

Woo, M. M., Chu, S. K. W., & Li, X. (2013). Peer-feedback and revision process in a wiki-mediated collaborative writing. *Educational Technology Research and Development*, 61(2), 279–309. https://doi.org/10.1007/s11423-012-9285-y

Zhang, C., & Magerko, B. (2025). *Generative AI Literacy: A Comprehensive framework for literacy and responsible use*. arXiv.org. https://arxiv.org/abs/2504.19038

Sezer Kizilates (sezerkzlates@gmail.com) is a CELTA-certified secondary school English teacher in Hong Kong. He holds a Master's degree in Technology, Design, and Leadership for Learning from HKU. His work focuses on technology-enhanced education, including the integration of generative AI tools into classroom practice and the use of innovative methods to boost student engagement and learning. He has received multiple awards in the Outstanding e-Learning Awards, including the Artificial Intelligence Special Award, and regularly shares his work at educational seminars, conferences, and professional development events.

Research Papers in Language Teaching and Learning
Vol. 16, No. 1, March 2026, 96-105
ISSN: 1792-1244
Available online at http://rpltl.eap.gr

Generative Artificial Intelligence in Language Education: Ethical Dimensions and the Equitable AI in Language Education Model

Eirini Ioanna Delmadorou

Generative Artificial Intelligence (GenAI) is reshaping language education by expanding opportunities for personalized feedback, materials development, and learner support. At the same time, its pedagogical promise is accompanied by significant ethical concerns involving bias, equity, transparency, and inclusivity. This paper critically examines these concerns through the lenses of sociocultural theory, second language acquisition (SLA), and constructivism, treating GenAI as a powerful but non-neutral cultural tool. The review highlights both the benefits of GenAI, such as increased access to adaptive input and output opportunities, and its risks, including algorithmic bias, Anglocentric dominance, and digital inequality. Building on this analysis, the paper introduces the *Equitable AI in Language Education Model* (EALEM), a conceptual framework organized around four guiding principles: inclusivity, transparency, human-in-the-loop mediation, and participatory design. EALEM offers a practical map for integrating GenAI in ways that support linguistic diversity, foster intercultural competence, and promote educational justice. By articulating this framework, the paper contributes to current debates on AI in education and proposes theoretically grounded, pedagogically relevant guidance for the responsible and equitable use of GenAI in language-learning contexts.

Keywords: GenAI, Language education ethics, Algorithmic bias, Digital equity, Sociocultural mediation, AI literacy

---

1. Introduction

The emergence of Generative Artificial Intelligence (GenAI) represents one of the most consequential technological developments in contemporary education. In recent years, tools such as large language models have begun to transform how learners engage with texts, how educators design instructional materials, and how institutions conceptualize teaching and learning (Li, 2025; Moorhouse & Wong, 2025). Within language education, GenAI is often presented as a catalyst for pedagogical innovation because it

can generate practice materials, provide rapid feedback, and create expanded opportunities for communicative rehearsal (Lee et.al,, 2026; Weng & Fu, 2025). Yet these pedagogical possibilities are inseparable from ethical questions concerning bias, equity, transparency, and cultural representation.

The ethical challenges associated with GenAI extend far beyond familiar concerns about plagiarism or academic misconduct. Scholars have drawn attention to algorithmic bias, digital inequality, and the reinforcement of Anglocentric norms in AI-generated materials (García- López et al., 2025; Gabriel, 2024). These concerns are particularly significant in language education, where linguistic identity, cultural meaning, and communicative legitimacy are central pedagogical issues rather than peripheral ones (Canagarajah, 2012; Pennycook, 2017). When generative systems rely disproportionately on English-dominant training data, they risk reproducing linguistic hierarchies and marginalizing less represented languages, thereby intensifying existing global inequalities (Nyaaba et al., 2024). At the same time, unequal access to advanced AI tools can deepen the digital divide between well-resourced and under-resourced educational settings (OECD, 2023; UNESCO, 2023).

The urgency of these issues is practical as well as theoretical. Educators are increasingly encountering GenAI in their classrooms and must decide how to integrate, regulate, or resist its use (Barnett, 2025; Stokes, 2025). The central challenge is how to balance the efficiency and creativity promised by these tools with the responsibility to cultivate critical thinking, inclusivity, and learner autonomy. This tension raises a series of questions: How can AI-generated materials represent diverse perspectives responsibly? Who is accountable when biased outputs shape learner experiences? How should teacher education evolve in order to prepare practitioners for the ethical dilemmas of AI-mediated classrooms?

Despite the growing volume of scholarship on AI in education, much of the literature remains focused on functionality rather than ethics. Systematic reviews have documented a rapid increase in empirical studies examining the use of GenAI tools, but comparatively fewer studies address fairness, transparency, or equity in educational practice (Chaudhry et al., 2022; Yan et al, 2025). Likewise, policy documents provide important high-level principles, yet they often stop short of offering pedagogically specific guidance for classroom implementation (U.S. Department of Education, 2023; UNESCO, 2024). There is, therefore, a clear need for theoretical contributions that identify ethical issues specific to language education and translate them into usable pedagogical guidance.

Although international frameworks increasingly articulate principles such as fairness, accountability, and transparency, these frameworks generally operate at the policy or systems level. Language education, however, faces distinctive ethical challenges because of its close relationship to culture, identity, power, and linguistic representation. This paper argues that existing AI ethics frameworks do not sufficiently address these language-specific dimensions and therefore do not fully support educators in operationalizing ethical AI use in language classrooms.

To address this gap, the present study introduces the *Equitable AI in Language Education Model* (EALEM), a pedagogically grounded framework that integrates sociocultural theory, SLA, and constructivist perspectives in order to provide classroom-oriented guidance for the ethical use of GenAI in language education. The paper critically examines the ethical dimensions of GenAI, identifies both risks and opportunities, and situates them within broader debates on linguistic justice and digital colonialism. It then proposes EALEM as a conceptual framework for responsible practice built around inclusivity, transparency, human-in-the-loop mediation, and participatory design. In doing so, the article aims to contribute not only to theoretical debate but also to practical, ethically grounded innovation in language education.

2. Theoretical Framework

Sociocultural theory offers a productive lens for understanding the integration of GenAI into language education. Rooted in the work of Vygotsky (1978), sociocultural perspectives conceptualize learning as a socially mediated process in which knowledge is constructed through interaction, guided participation, and the use of cultural tools. From this perspective, GenAI can be understood as a mediational artifact that shapes how learners access, produce, and negotiate language. AI-powered prompts, corrective feedback, and dialogic simulations may extend the learner's zone of proximal development by scaffolding performance in ways broadly consistent with Vygotskian principles (Lantolf & Thorne, 2006). At the same time, sociocultural theory cautions against treating tools as neutral: the assumptions, values, and biases embedded in GenAI systems inevitably influence the learning experiences they mediate.

Complementing this perspective are theories of second language acquisition, particularly Krashen's Input Hypothesis and Swain's Output Hypothesis. Krashen (1982) emphasized the importance of comprehensible input as a driver of acquisition, whereas Swain (1995) highlighted the value of pushed output in consolidating linguistic knowledge. GenAI has the potential to influence both processes. On the one hand, large language models can provide learners with abundant, adaptable input through generated dialogues, explanations, and simplified texts. On the other hand, interactive AI systems can create low-stakes opportunities for language production by prompting learners to write, speak, revise, and reformulate their responses (Moorhouse & Wong, 2025). Yet, as Ellis (2003) suggests, interaction in SLA is not merely the exchange of forms; it is also shaped by negotiation, reciprocity, and meaning-making. For that reason, GenAI should be designed to complement rather than replace interaction with teachers and peers.

Constructivist theories further illuminate the pedagogical significance of GenAI in language education. Constructivism emphasizes the active role of learners in building knowledge through exploration, interpretation, and collaboration. From this perspective, AI tools may support learning when they enable learners to test ideas, receive feedback, and revise their understanding in iterative ways. However, constructivism also reminds us that meaningful learning depends on reflection and agency rather than on passive reception of ready-made answers. The educational value of GenAI, therefore, depends not simply on access to technological outputs but on the quality of the pedagogical tasks in which those outputs are embedded.

To connect these theoretical perspectives to the ethical focus of the paper, four core concepts guide the present analysis: equity, fairness, inclusivity, and transparency. Equity refers to the just distribution of opportunities and support in ways that recognize structural differences in learners' needs and circumstances. Fairness concerns the avoidance of discriminatory outcomes and the responsible treatment of learners across linguistic, cultural, and social groups. Inclusivity involves designing learning environments and resources that reflect diversity and allow meaningful participation. Transparency concerns the intelligibility of AI systems, including the extent to which their purposes, limitations, and decision-making processes can be made understandable to educators and learners. These concepts are interdependent: transparency supports fairness by making bias more visible, and inclusivity strengthens equity by ensuring that historically marginalized learners are not excluded from the benefits of innovation.

By integrating sociocultural theory, SLA, constructivism, and these ethical concepts, this paper situates GenAI within a coherent theoretical framework. Sociocultural theory highlights its status as a cultural tool; SLA clarifies its implications for input, output, and interaction; constructivism underscores learner agency

and reflective engagement; and the ethical concepts of equity, fairness, inclusivity, and transparency provide the normative criteria through which AI integration in language education should be evaluated.

3. Literature Review

3.1 Applications of Generative Artificial Intelligence in Language Education

The integration of GenAI into language education has expanded rapidly, with scholars documenting a wide range of applications that affect both instructional practice and learner experience. One of the most prominent uses of GenAI is the creation of instructional materials. Educators can employ AI tools to generate reading passages, vocabulary activities, grammar exercises, discussion prompts, and differentiated materials tailored to diverse proficiency levels (Lee et al., 2026; Moorhouse & Wong, 2025). This flexibility is particularly attractive in language education, where teachers often need to adapt content to different learner needs, interests, and contexts.

GenAI also supports feedback and formative assessment. Compared with earlier forms of computer-assisted language learning, generative models can provide more nuanced explanations, examples, and reformulations that approximate dialogic feedback. Such tools may help learners notice errors, experiment with language choices, and revise their texts in real time. This can increase opportunities for practice, especially in settings where teachers face large classes or limited time (Moorhouse & Wong, 2025; Weng & Fu, 2025).

A growing body of scholarship situates these applications within broader processes of educational digitalization. Li (2025), in a systematic review of empirical research on GenAI in education, notes that many studies focus on efficiency, personalization, and learner engagement. In language education specifically, GenAI is often praised for lowering barriers to practice by enabling conversational simulation, vocabulary support, and immediate feedback. These developments suggest that GenAI may contribute to more adaptive and responsive language learning environments. However, the same literature also makes clear that pedagogical usefulness alone cannot serve as the sole criterion for adoption; ethical considerations remain central.

3.2 Ethical Concerns

Although the pedagogical potential of GenAI is substantial, its widespread adoption raises a series of ethical concerns that require critical scrutiny. Among the most pressing is algorithmic bias. García-López et al. (2025) argue that GenAI systems can reproduce social and cultural inequalities because they are trained on data that reflect existing imbalances in representation and power. In language education, such bias may appear in stereotypical cultural portrayals, the privileging of dominant varieties of English, or the underrepresentation of minority languages and communicative norms. Gabriel (2024) similarly warns that educational uses of GenAI may widen inequity unless issues of access, representation, and cultural responsiveness are addressed directly.

Transparency and accountability constitute a second major concern. Chaudhry et al. (2022) propose a transparency framework for AI in education and emphasize the importance of making system logic, limitations, and outputs understandable to educators and learners. Yet many widely used GenAI tools operate as opaque “black boxes,” making it difficult to determine how outputs are generated, which sources of bias are embedded in the system, or how errors should be interpreted. Dwivedi et al. (2023)

likewise stress that institutions often prioritize innovation and efficiency more readily than sustained ethical reflection, thereby creating governance gaps in educational AI adoption.

Equity and access form a third area of concern. OECD (2023) and UNESCO (2024) both underline the risk that AI adoption may reinforce existing educational inequalities if access to quality tools, infrastructure, and teacher training remains uneven. In language education, this risk is especially acute because AI tools can become gatekeepers to high-quality input, feedback, and learning support. Well-resourced institutions may be able to integrate these tools effectively, while under-resourced contexts remain excluded from their potential benefits.

Finally, ethical concerns extend to learner autonomy. While GenAI can support learners by offering immediate assistance, explanations, and practice opportunities, it may also foster overreliance if used uncritically. Giannakos (2024) notes that the promise of efficiency can obscure deeper questions about what learners are actually doing cognitively when they rely on AI-generated responses. If learners begin to outsource idea generation, revision, and reflection too readily, GenAI may weaken rather than strengthen the independent and critical capacities that language education seeks to cultivate.

3.3 Policy and Guidelines

The growing awareness of these ethical challenges has led to the development of policy frameworks and institutional guidance. UNESCO (2024) has issued recommendations for the use of generative AI in education and research, highlighting the importance of transparency, accessibility, data governance, and human oversight. Similarly, the U.S. Department of Education (2023) frames AI as both an opportunity and a responsibility, arguing that innovation must be accompanied by protections for equity, safety, and educational quality.

National and institutional responses are also evolving. Cassidy (2024) reports that Australian schools have moved toward more formal guidance on classroom AI use, reflecting an attempt to balance opportunity with risk. Barnett (2025) and Stokes (2025) likewise suggest that educators are increasingly experimenting with AI tools while simultaneously recognizing concerns about cheating, data privacy, bias, and professional preparedness. These accounts reveal that policy development is no longer abstract; it is becoming a practical issue of classroom governance and professional judgment.

At the same time, the literature suggests that policy cannot be limited to top-down regulation. García-López et al. (2025) argue that ethical AI in education requires participatory approaches in which educators and learners help shape how systems are implemented. This is especially important in language education, where local linguistic ecologies, cultural identities, and curricular goals vary substantially across contexts. Overall, the literature depicts a field marked by both enthusiasm and caution: GenAI offers real pedagogical affordances, but its ethical implications remain insufficiently resolved.

4. Discussion

The literature reviewed above indicates that the integration of GenAI into language education cannot be reduced to a simple debate between innovation and resistance. Rather, it reveals a more complex terrain in which pedagogical opportunity and ethical risk are deeply intertwined. Language education is especially sensitive to these tensions because language learning is never merely technical; it involves identity,

culture, legitimacy, and participation in wider social worlds. As a result, the ethical evaluation of GenAI in this field must attend to more than functional effectiveness.

A first point of tension concerns representational bias. Large language models inherit the limitations of their training data and may reproduce dominant cultural assumptions in subtle but consequential ways (García-López et al., 2025; Yan et al., 2025). In language classrooms, such bias may shape which accents appear legitimate, which cultural narratives are treated as normative, and which communicative practices are rendered visible or invisible. From the perspective of translingual practice, this is particularly problematic because language education should expand rather than constrain learners' engagement with linguistic diversity (Canagarajah, 2012). A second issue concerns access and structural inequality. OECD (2023) and UNESCO (2024) show that digital inequality remains a major barrier to equitable AI integration. If high-quality AI tools become concentrated in privileged institutions, then GenAI may intensify rather than alleviate educational disparities. The problem is not only material access to devices and subscriptions, but also access to training, institutional support, and pedagogical guidance. In this sense, AI literacy is itself an equity issue. Third, the educator's role must be reconsidered. As GenAI systems assume some functions traditionally associated with teachers, such as generating materials, responding to learner questions, or suggesting revisions, the educator's work shifts from information delivery toward ethical mediation and pedagogical judgment. This does not diminish the teacher's importance; on the contrary, it increases the need for professional discernment. Educators must evaluate outputs, contextualize them, identify bias, and decide when AI use supports learning and when it undermines it (Richards & Rodgers, 2014; U.S. Department of Education, 2023).

Learner autonomy also requires careful reinterpretation. GenAI can create opportunities for independent practice and immediate support, which may increase confidence and engagement. Yet autonomy in language education does not mean simply working alone with a tool; it means developing the capacity to make informed, reflective, and critical choices. If learners treat AI output as authoritative or become dependent on automated assistance, the apparent independence offered by AI may conceal a loss of intellectual agency. Constructivist perspectives, therefore, suggest that GenAI should be used to stimulate inquiry and revision, not to replace them.

Finally, the literature raises broader concerns about digital colonialism. Nyaaba et al. (2024) argue that AI systems developed primarily in the Global North can export dominant epistemologies, languages, and values to other educational settings. In language education, this risk is especially serious because linguistic diversity is inseparable from cultural recognition and educational justice. The ethical challenge is not simply to add more languages to AI systems, but to ensure that local knowledge, communicative practices, and learner identities are treated as pedagogically meaningful rather than as peripheral deviations from a dominant norm. Taken together, these concerns do not justify rejecting GenAI outright. Rather, they point to the need for an explicit ethical framework capable of translating broad principles into pedagogically meaningful guidance. It is in response to that need that the present paper proposes the *Equitable AI in Language Education Model* (EALEM).

5. Proposed Conceptual Framework

Unlike Equitable *AI in Language Education Model* (EALEM), which is proposed as a language-specific, pedagogically grounded framework for the ethical integration of GenAI. Unlike general AI ethics frameworks that remain at the level of broad normative principles, EALEM is designed to address the particular realities of language education, where issues of identity, representation, interaction, and access are central. The model is grounded in the theoretical perspectives outlined earlier and organized around

four interrelated principles: inclusivity, transparency, human-in-the- loop mediation, and participatory design.

The first principle, *inclusivity*, refers to the intentional design and pedagogical use of GenAI tools in ways that reflect linguistic diversity, cultural plurality, and learner identity. An inclusive AI-mediated language classroom does not treat learners as a homogeneous group or present dominant language varieties as universally normative. Instead, it critically examines whose voices, examples, and communicative practices are represented in AI-generated content. In practical terms, this principle calls for educators to review outputs for cultural stereotyping, expand the range of linguistic examples used in instruction, and adapt AI tasks so that diverse learners can participate meaningfully.

The second principle, *transparency,* addresses the opacity of generative systems. Because many GenAI tools function as black boxes, educators and learners may not understand how outputs are produced or why particular responses appear authoritative (Chaudhry et al., 2022). Within EALEM, transparency involves making the limitations, risks, and provisional nature of AI output explicit. This means discussing hallucinations, bias, dataset limitations, and uncertainty with learners rather than presenting AI-generated text as neutral knowledge. Transparency also requires institutions to adopt tools and policies that permit meaningful scrutiny wherever possible.

The third principle, *human-in-the-loop mediation*, emphasizes that GenAI should support rather than displace human pedagogical judgment. Grounded in sociocultural and constructivist understandings of learning, this principle positions educators as mediators who interpret, contextualize, and evaluate AI use in relation to pedagogical goals. Human oversight is especially important in language education because feedback, correction, and cultural framing are not value-free processes. Under EALEM, teachers remain responsible for deciding when AI use is educationally appropriate, how outputs should be discussed, and how learner reflection can be sustained.

The fourth principle, *participatory design*, extends ethical reflection beyond classroom practice to the broader processes through which AI tools are selected, adapted, and governed. García- López et al. (2025) argue that ethical AI in education requires the involvement of those who are affected by its implementation. In the context of language education, this means that educators, learners, and communities should have a voice in how AI systems are integrated, evaluated, and refined. Participatory design is particularly important in multilingual and culturally diverse settings, where top-down technological solutions may overlook local educational priorities and linguistic realities.

Together, these four principles form a framework that is both theoretically grounded and practically adaptable. Inclusivity ensures representation and cultural responsiveness; transparency supports accountability and critical understanding; human-in-the-loop mediation preserves pedagogical judgment; and participatory design promotes legitimacy, responsiveness, and context sensitivity. EALEM does not offer a universal formula, but it provides a structured way of thinking about ethical integration that can guide educators, institutions, and researchers. The implications of EALEM extend beyond individual classrooms. In teacher education, the framework suggests the need to prepare educators not only in technical AI literacy but also in ethical reasoning, bias recognition, and reflective task design. In institutional policy, it points toward more context-sensitive forms of governance that connect general principles to local pedagogical realities. In research, it offers a conceptual basis for examining how ethical AI integration can be operationalized and evaluated across different language-learning contexts.

EALEM also has limitations. Implementing inclusivity may be constrained by the availability of culturally and linguistically diverse datasets. Transparency is difficult when commercial systems do not disclose their underlying processes. Human-in-the-loop mediation requires time, expertise, and institutional support that may not always be available. Participatory design, while desirable, depends on genuine collaboration between educators, learners, developers, and policymakers. For these reasons, EALEM should be understood not as a final solution but as a starting point for continued dialogue, critical reflection, and empirical refinement.

6. Conclusion

The integration of GenAI into language education is both an opportunity and an ethical test. This paper has argued that although GenAI can support personalization, feedback, and expanded access to language practice, its adoption also raises serious questions about bias, equity, opacity, and learner autonomy (Lee et al., 2026; Moorhouse & Wong, 2025). Through the lenses of sociocultural theory, SLA, and constructivism, the discussion has shown that GenAI functions as a mediational tool that reshapes input, output, interaction, and the conditions under which language learning takes place (Krashen, 1982; Swain, 1995; Vygotsky, 1978).

The proposed *Equitable AI in Language Education Model* (EALEM) responds to these challenges by articulating four guiding principles: inclusivity, transparency, human-in-the-loop mediation, and participatory design. Taken together, these principles provide a conceptual map for the ethically grounded adoption of GenAI in language education. They also reframe the role of educators as ethical mediators who must help learners engage critically with AI-generated content rather than consume it unreflectively (Richards & Rodgers, 2014; UNESCO, 2024).

The broader implications of this argument are pedagogical, institutional, and political. Pedagogically, language classrooms must cultivate critical AI literacy alongside linguistic competence. Institutionally, schools and universities must invest in professional learning, infrastructure, and policy guidance that support equitable implementation. Politically, policymakers must recognize that AI in education is not only a matter of innovation, but also of justice. Without deliberate safeguards, GenAI may reproduce digital colonialism and deepen global inequities (Nyaaba et al., 2024). With thoughtful, ethically informed use, however, it may support more inclusive and responsive language-learning environments.

Future research should test EALEM in diverse educational contexts, including multilingual classrooms and under-resourced settings, in order to examine its feasibility, limitations, and pedagogical impact. Such research should move beyond the question of what AI can do and address the more demanding question of what AI should do in education (Dwivedi et al., 2023). The future of GenAI in language education will depend less on the technology itself than on the values, frameworks, and practices that shape its use. If approached critically and ethically, GenAI may not become a substitute for human teaching but a tool that supports educational justice, linguistic diversity, and reflective learning.

References

Barnett, S. (2025, August 18). Teachers are trying to make AI work for them. *Wired.* https://www.wired.com/story/teachers-using-ai-schools

Canagarajah, S. (2012). Translingual Practice: Global Englishes and Cosmopolitan Relations (1st ed.). Routledge. https://doi.org/10.4324/9780203073889

Chaudhry, A., Cukurova, M., & Luckin, R. (2022). *A transparency index framework for AI in education*. arXiv. https://arxiv.org/abs/2206.03220

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., ... & Wright, R. (2023). Opinion Paper: "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,* 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642

Ellis, R. (2003). *Task-based language learning and teaching*. Oxford University Press.

Gabriel, S. (2024). Generative AI and educational (in)equity. *Proceedings of the International Conference on AI Research, 4*(1), 133-142. https://doi.org/10.34190/icair.4.1.3153.

García-López, I. M., Sánchez-Vera, M. M., Solano-Fernández, I. M., & Pérez-Hernández, D. (2025). Ethical and regulatory challenges of generative AI in education. *Frontiers in Education*, 10, 1565938. https://doi.org/10.3389/feduc.2025.1565938

Giannakos, M. (2024). The promise and challenges of generative AI in education. *Behaviour & Information Technology*, *43*(11), 1299–1312. https://doi.org/10.1080/0144929X.2024.2394886

Krashen, S. D. (1982). Principles and practice in second language acquisition. Pergamon Press.

Lantolf, J. P., & Thorne, S. L. (2006). Sociocultural theory and the genesis of second language development. Oxford University Press.

Lee, S., Choe, H., Zou, D., & Jeon, J. (2026). Generative AI (GenAI) in the language classroom: A systematic review. *Interactive Learning Environments, 34(1)*, 335–359. https://doi.org/10.1080/10494820.2025.2498537

Li, B. (2025). Two years of innovation: A systematic review of empirical generative AI research in language learning and teaching. *Computers & Education: Artificial Intelligence, 8,* 100445. https://doi.org/10.1016/j.caeai.2025.100445

Moorhouse, B. L. (2025). *Generative artificial intelligence and language teaching*. Cambridge University Press. https://doi.org/10.1017/9781009618823

Nyaaba, M., Wright, A., & Choi, G. L. (2024). *Generative AI and digital neocolonialism in global education*. https://doi.org/10.48550/arXiv.2406.0296

OECD. (2023). *AI in education: Ensuring equity and transparency*. OECD Publishing. https://doi.org/10.1787/ai-education-en

Pennycook, A. (2017). *Posthumanist applied linguistics*. Routledge.

Richards, J. C., & Rodgers, T. S. (2014). *Approaches and methods in language teaching* (3rd ed.). Cambridge University Press.

Stokes, K. (2025, September 2). Minnesota schools embrace AI—cautiously. *Axios Twin Cities.* https://www.axios.com/local/twin-cities/2025/09/02/ai-back-to-school-minnesota

Swain, M. (1995). Three functions of output in second language learning. In G. Cook & B. Seidlhofer (Eds.), *Principle and practice in applied linguistics* (pp. 125–144). Oxford University Press.

U.S. Department of Education. (2023). *Artificial intelligence and the future of teaching and learning*. https://files.eric.ed.gov/fulltext/ED631097.pdf

UNESCO. (2023). *Guidance for generative AI in education and research*. UNESCO Publishing. https://unesdoc.unesco.org/ark:/48223/pf0000386693

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Weng, Z., & Fu, Y. (2025). Generative AI in language education: Bridging divide and fostering inclusivity. *International Journal of Technology in Education, 8*(2), 395-420. https://doi.org/10.46328/ijte.1056

Yan, Y., Liu, H., & Chau, T. (2025). A Systematic Review of AI Ethics in Education: Challenges, Policy Gaps, and Future Directions. Journal of Global Information Management (JGIM), 33(1), 1-50. https://doi.org/10.4018/JGIM.386381

Eirini Ioanna Delmadorou (reniadelmar@gmail.com) is a philologist, English language educator, and certified adult trainer. She holds a BA in Philosophy from the National and Kapodistrian University of Athens and an MA in Teaching English as a Foreign/International Language from the Hellenic Open University. Her academic background includes specialisation in Political and Economic Philosophy, Bioethics, and the intersection of Philosophy and Technology. She has extensive teaching experience in secondary education and private tutoring, teaching Ancient and Modern Greek, History, Philosophy, Latin, and English as a Foreign Language. She has also completed teaching placements at the Model High School of the Ionideios School of Piraeus.

Research Papers in Language Teaching and Learning
Vol. 16, No. 1, March 2026, 106-140
ISSN: 1792-1244
Available online at http://rpltl.eap.gr

Integrating Large Language Models into Corpus-Based Teaching: A Framework for Speech-Language Pathology, Computational Linguistics and Clinical Programs

Athanasios Karasimos, Evangelia-Antonia Efstratiadou, Christos Papatzalas & Ilias Papathanasiou

This paper proposes a systematic framework for integrating Large Language Models (LLMs) into corpus-based teaching for Speech-Language Pathology, Computational Linguistics, and Clinical Neurolinguistics programs. Traditional corpus analysis, though essential, presents significant pedagogical challenges, including being time-intensive and requiring specialized technical expertise, especially with atypical language data like aphasic speech. The emergence of LLMs offers a transformative opportunity to overcome these barriers. The systematic framework introduces practical educational modules, validated annotation workflows, and assessment strategies, exemplified through the Greek CACLA corpus. Crucially, the approach advocates for using LLMs as "cognitive partners" to handle routine tasks, allowing students to focus on higher-order analysis, clinical interpretation, and the critical evaluation of AI outputs. This method aims to democratize sophisticated linguistic analysis while ensuring students develop necessary critical thinking capacities and technological literacy.

Keywords: Large Language Models, Computational Linguistics, Speech-Language Pathology, Clinical Neurology, Aphasia, Corpus Methodology

---

1. Introduction

The integration of corpus linguistics into higher education has long been recognized as essential for training students in language sciences, such as speech-language pathology, computational linguistics, and clinical neuroscience programs (Reppen, 2010; Römer, 2011). However, traditional corpus analysis methods present significant pedagogical challenges: they are time-intensive, require specialized technical expertise, and often create barriers between linguistic theory and clinical

application (Armstrong, 2000; Friginal & Hardy, 2014). Educators and students frequently struggle with the challenging learning curve of annotation software, and the difficulty of extracting meaningful patterns from clinical language data (Ball, 2013; Crystal, 2013). These challenges are particularly delicate when working with atypical language corpora, such as aphasic speech, where linguistic irregularities complicate automated annotations and analysis, and demand advanced interpretive skills (Goodglass *et al*., 2001).

The emergence of Large Language Models (LLMs), such as GPT-5, Gemini and Claude among others, represents a transformative opportunity for corpus-based education (Brown *et al*., 2020). These advanced AI systems demonstrate significant and evolving capabilities in linguistic annotation, pattern recognition, but still with several issues on natural language understanding (Bommasani *et al*., 2021; Zhao *et al*., 2025). Recent studies explore LLM applications in linguistic research, including morphological analysis (Mita *et al*., 2024; Karasimos & Makri, 2025; Karasimos, Makri & Petropoulou, 2025), syntactic parsing (Khan, 2025; Manning, 2022); yet their pedagogical integration remains largely unexplored, particularly in clinical and computational linguistics education or LLMs have been (extensively) mis-presented and ill-researched by non-AI-expertised educators.

Despite growing theoretical and research interest in AI-enhanced education (Holmes *et al*., 2019; Zawacki-Richter *et al*., 2019) and recognition of technology's role in language teaching and speech-language pathology training (Keck & Doarn, 2014; Hill *et al*., 2020), a significant research gap still remains; no comprehensive framework currently guides educators in integrating LLMs, corpora and educational technology tools into their curricula. This paper addresses this gap by presenting a systematic framework for incorporating LLMs and corpus-based teaching across speech-language pathology, computational linguistics, and clinical neurolinguistics courses. We introduce practical educational modules, validated annotation workflows, and assessment strategies, exemplified through the CACLA (Corpus for Aphasic Clinical Language Analysis) Greek corpus. Our aim is to provide educators, teachers, and practitioners with tried-and-tested tools that embed LLM capabilities while maintaining pedagogical aspect, critical thinking development, and awareness of technological limitations.

2. From Corpora to Classrooms: Bridging Clinical Language Data and Educational Practice

The path from analyzing aphasic speech corpora to developing effective educational frameworks requires understanding both the clinical linguistic landscape and the pedagogical challenges inherent in teaching corpus analysis methods. This section traces the evolution from traditional corpus analysis in communication disorders to the emerging potential of Large Language Models, while situating these developments within the broader context of technology-enhanced education in speech-language pathology and computational linguistics.

2.1 Corpus Analysis in Communication Disorders

Corpus-based approaches to studying communication disorders have transformed the understanding of atypical language patterns over the past three decades (MacWhinney *et al*., 2011; Ball *et al*., 2008). The systematic collection and analysis of language data from individuals with aphasia, language disorders, and other communication impairments offered the opportunity to identify quantifiable linguistic markers, patterns of recovery trajectories, and validate assessment protocols/tests (Armstrong, 2000; Berndt *et al*., 2000; Saffran *et al*., 1989; Wilson *et al*., 2017).

Clinical corpora serve multiple critical functions mostly in research (and less in education). They provide authentic data that capture the variability and complexity of disordered language production

(Prins & Bastiaanse, 2004; Stark *et al*., 2025; Varlokosta *et al*., 2016; Webster *et al*., 2007); for instance, the AphasiaBank corpus[1], contains hundreds of discourse samples from people with aphasia across diverse tasks and several languages, providing access to several phenomena ranging from word retrieval difficulties, error samples to discourse coherence (Dalton & Richardson, 2019; MacWhinney *et al*., 2011). Similarly, specialized corpora like the Northwestern Narrative Language Analysis (NNLA)[2] system provide structured protocols for analyzing story retelling, and offering insights into significant content and linguistic efficiency (Thompson *et al*., 2012; Kintz & Wright, 2018).

Based on the Greek digital landscape, corpus resources for communication disorders remain relatively limited, though several significant initiatives have emerged, such THALIS Aphasia (Varlokosta *et al*., 2016; Βαρλοκώστα *et al*., 2017) and some smaller individual research projects (Kambanaros & van Steenbrugge, 2006; Nanousi *et al*., 2006). The CACLA corpus represents a comprehensive effort to document Greek spoken and aphasic speech across various aphasia types and task conditions, providing rich material for both clinical research and educational applications (Papathanasiou *et al*., 2013).

The linguistic analysis of aphasic corpora presents unique challenges that distinguish it from typical language corpus research. Atypical language production includes phenomena such as *agrammatism* (simplified syntax with function word omissions), *paragrammatism* (syntactically complex but semantically empty output), *phonemic and semantic paraphasias* (sound-based and meaning-based word substitutions), *neologisms* (invented non-words or non-sense words), and discourse-level disruptions including tangentiality and reduced coherence (Caplan, 1987; Goodglass *et al*., 2001; Kolk, 1995). These features require experienced annotators to make complex linguistic judgments, distinguishing between intentional linguistic choices and pathological productions—a distinction that challenges both human and automated annotation systems (Rochon *et al*., 2000; Bastiaanse & Edwards, 2004).

Educational applications of clinical corpora have primarily focused on introducing students to authentic clinical data, teaching linguistic analysis skills, and illustrating theoretical concepts with real-world examples (Barbieri & Eckhardt, 2007; Römer, 2011). However, the labor-intensive nature of corpus annotation and analysis has limited widespread adoption in educational settings, particularly in programs lacking computational linguistics infrastructure (Friginal & Hardy, 2014; Reppen, 2010).

2.2 Large Language Models for (Spoken Data) Text Analysis

Recent research has begun exploring LLM applications across linguistic domains. Morphological analysis studies demonstrate high accuracy for part-of-speech tagging and morphological feature extraction across diverse languages, including morphologically complex ones (Liu *et al*., 2024; Üstün *et al*., 2024). For Greek specifically, LLMs show promise in handling rich inflectional morphology, though systematic evaluation remains limited (Koutsikakis *et al*., 2020).

Syntactic parsing evaluations reveal that LLMs can produce dependency and constituency parses approaching state-of-the-art specialized parsers, with particular strength in handling ambiguous structures through contextual reasoning (Kulmizev *et al*., 2024; Warstadt & Bowman, 2022). However, performance varies with sentence complexity and grammaticality—a critical consideration for clinical corpus analysis (Hu *et al*., 2020).

[1] https://talkbank.org/aphasia/

[2] https://talkbank.org/aphasia/discourse/C-NNLA/

Semantic analysis applications include word sense disambiguation, semantic role labeling, and metaphor detection, with LLMs often surpassing traditional systems by leveraging vast world knowledge encoded during pre-training (Pilehvar & Camacho-Collados, 2019; Ettinger, 2020). Discourse analysis capabilities extend to coreference resolution, coherence evaluation, and rhetorical structure identification, though systematic evaluation frameworks remain under development (Hao *et al*., 2021).

Error detection and correction represent particularly relevant applications for clinical linguistics. Studies show LLMs can identify grammatical errors, semantic anomalies, and stylistic inconsistencies, suggesting potential for automated coding of aphasic errors (Shen *et al*., 2023; Wu *et al*., 2023). However, the ability to distinguish pathological from intentional non-standard forms—crucial for clinical analysis—requires careful prompt engineering and validation (Bryant *et al*., 2023).

Cross-linguistic applications demonstrate that multilingual LLMs like GPT-5 and mBERT can perform reasonably well on lower-resourced languages, though performance typically degrades for languages with limited training data representation (Joshi *et al*., 2020; Üstün *et al*., 2024). For Greek, intermediate resource availability suggests moderate but not optimal LLM performance without fine-tuning (*Koutsikakis* et al., 2020).

The application of LLMs to atypical or clinical language data represents a relatively unexplored frontier. Preliminary investigations suggest both promise and challenges. LLMs demonstrate robustness to disfluency, maintaining reasonable parsing accuracy even with filled pauses, false starts, and self-corrections common in aphasic speech (Bhat *et al*., 2022). However, handling of agrammatism—telegraphic speech with systematic function word omissions—proves more challenging, as LLMs may inappropriately "correct" such productions rather than analyzing them as-produced.

Paraphasia detection studies (Zhu *et al*., 2023) suggest LLMs can identify semantic substitutions when provided with appropriate context and task framing, though phonemic paraphasias may require explicit phonological representation. Neologism identification presents particular difficulty, as LLMs may misinterpret non-words as rare real words, code-switching or hapax legomena, requiring explicit instruction to flag potential neologisms (Kuzman *et al*., 2023). Discourse-level analysis of clinical narratives shows promise for automated coherence rating and information content scoring. Studies applying GPT-based models to aphasia discourse samples report moderate correlations with human ratings of narrative quality (Fergadiotis *et al*., 2023). However, LLMs may over-emphasize semantic coherence while missing subtle linguistic markers of impairment, necessitating careful validation against clinical gold standards. Variability and consistency concerns emerge prominently in clinical applications. LLMs can produce different outputs for identical inputs across sampling runs, problematic for research requiring precise replicability (Ouyang *et al*., 2022). Temperature settings, random seeds, and prompt variations all influence output consistency—factors requiring systematic control in educational and research contexts (Sclar *et al*., 2023).

2.3 Technology in SLP Education

The integration of technology into speech-language pathology education has evolved substantially over recent decades, though adoption patterns vary considerably across institutions and geographic regions (Hill *et al*., 2020). Simulation and virtual patients represent well-established technologies in SLP education, allowing students to practice clinical interactions, assessment procedures, and intervention planning in risk-free environments (Dudding et al., 2011; Hill & Theodoros, 2017).

Platforms like SimuCase[3] provide standardized virtual cases covering diverse communication disorders, enabling repetitive practice and immediate feedback. However, these systems typically present scripted scenarios rather than authentic language samples, limiting exposure to real-world linguistic variability.

Video analysis tools enable students to review authentic clinical interactions, annotate communicative behaviors, and analyze language samples from recorded sessions. Software like ELAN[4] allows time-aligned transcription and annotation, though the labor-intensive nature limits typical use to small sample sizes (Sloetjes & Wittenburg, 2008). Computer-assisted learning modules deliver instructional content on anatomical structures, phonetic transcription, and assessment procedures through interactive multimedia (Brackenbury et al., 2008). While valuable for knowledge transmission, these typically lack integration with authentic clinical data analysis. Telepractice platforms facilitate remote clinical supervision, enabling students to observe real therapy sessions and receive feedback from clinical educators regardless of geographic location (Grogan-Johnson *et al*., 2013; Tucker, 2012). The COVID-19 pandemic accelerated adoption of telehealth modalities, demonstrating feasibility while highlighting technological barriers and training needs.

Specialized software for language analysis, including systematic analysis of language transcripts (SALT) and CLAN, provides quantitative analysis capabilities for clinical language samples (MacWhinney, 2000; Miller & Iglesias, 2020). However, educational surveys indicate limited integration into graduate curricula, attributed to perceived complexity, time investment, and insufficient instructor familiarity (Brackenbury *et al*., 2008). Broader healthcare education has increasingly incorporated artificial intelligence applications, providing relevant models for SLP education. Medical image analysis using deep learning algorithms helps radiology students develop diagnostic skills through immediate feedback on interpretation accuracy (Hosny *et al*., 2018; Shen *et al*., 2017). Clinical decision support systems powered by machine learning assist medical students in differential diagnosis, treatment planning, and predicting patient outcomes (Beam & Kohane, 2018; Rajkomar *et al*., 2019). Natural language processing applications in medical education include automated analysis of clinical notes, extraction of relevant information from electronic health records, and evaluation of documentation quality (Koleck *et al*., 2019; Wang *et al*., 2018). These applications demonstrate feasibility of integrating sophisticated AI tools into clinical training while maintaining pedagogical rigor. Intelligent tutoring systems adapt instructional content and feedback based on individual student performance, demonstrating learning gains compared to traditional instruction across medical disciplines (Kulik & Fletcher, 2016; Woolf *et al*., 2013). However, concerns regarding over-reliance on automation, reduced critical thinking, and "black box" decision-making have prompted calls for careful pedagogical design emphasizing human oversight and transparent reasoning (Holmes *et al*., 2019; Zawacki-Richter *et al*., 2019).

3. The educational prism of AI, LLMs and corpora in the tertiary system

The integration of artificial intelligence, Large Language Models, and corpus linguistics into tertiary education represents a convergence of technological innovation and pedagogical necessity. This section examines the theoretical foundations that underpin effective educational implementation, drawing from established learning frameworks, technology integration models, and corpus-based pedagogical research. We focus specifically on how these theoretical perspectives apply to computational linguistics, neurolinguistics, and speech-language pathology education—disciplines where authentic language data, analytical rigor, and clinical competence intersect.

[3] https://www.simucase.com/

[4] https://archive.mpi.nl/tla/elan

3.1 Educational Framework: Learning Objectives and Cognitive Development

The development of effective learning objectives for LLM-based corpus analysis requires careful alignment with program-specific competencies while addressing the unique cognitive demands of working with both clinical language data and artificial intelligence tools. Learning objectives must balance technical skill acquisition with critical evaluation capabilities, recognizing that students need not only to operate AI systems but also to understand their limitations and validate their outputs (Anderson & Krathwohl, 2001; Bloom et al., 1956; Fink, 2013).

For computational linguistics programs, learning objectives emphasize algorithmic understanding, evaluation methodology, and the ability to critically assess automated linguistic analysis. Students should be able to design annotation schemes, implement validation procedures, calculate inter-annotator agreement metrics, and compare LLM performance against traditional NLP approaches (Liddy, 2001; Jurafsky & Martin, 2023). These objectives align with computational thinking frameworks that emphasize decomposition of complex problems, pattern recognition, abstraction, and algorithmic design (Weintrop *et al*., 2016; Wing, 2006).

For speech-language pathology programs, learning objectives center on clinical application: identifying linguistically-based assessment measures, interpreting quantitative indices of language impairment, connecting corpus patterns to theoretical models of aphasia, and making evidence-based clinical decisions informed by linguistic analysis (American Speech-Language-Hearing Association [ASHA], 2016). The integration of corpus analysis with clinical reasoning represents what Fink (2013) terms "integration learning"—the ability to connect ideas, perspectives, and realms of life, essential for translating linguistic research into clinical practice.

For neurolinguistics and clinical neuroscience programs, objectives emphasize understanding brain-language relationships through systematic analysis of language breakdown patterns. Students must develop competencies in identifying neuroanatomical correlates of linguistic deficits, recognizing dissociations between preserved and impaired language functions, and critically evaluating neurolinguistic theories based on corpus evidence (Hillis, 2007; Papathanasiou *et al*., 2017).

3.2 Technology Integration in Teaching and Learning: Frameworks and Models TPACK Framework for AI-Enhanced Corpus Linguistics Education

The TPACK framework (Mishra & Koehler, 2006) guides integration of LLM-based corpus analysis into linguistics education, recognizing that effective technology use requires interplay between Technological (TK), Pedagogical (PK), and Content Knowledge (CK). TK encompasses LLM capabilities, prompt engineering, and computational principles underlying language models. PK involves scaffolding, assessment, and fostering critical thinking, while CK includes linguistic theory, clinical frameworks, and corpus methodology. The intersections are crucial, since TCK means understanding how AI tools address specific linguistic phenomena; PCK involves effective teaching strategies for complex concepts; TPK recognizes how LLMs enable students to analyze larger datasets, revealing patterns previously accessible only to researchers. Full TPACK integration—e.g., teaching agrammatism through collaborative LLM-assisted annotation that students validate—transforms not just what students learn but how (Koehler & Mishra, 2009).

The SAMR model (Puentedura, 2006) evaluates transformative potential across four levels: Substitution (LLMs replace manual annotation with no pedagogical change), Augmentation (students annotate larger datasets with immediate feedback), Modification (students compare LLM performance across aphasia types, integrating linguistic and computational thinking), and Redefinition

(real-time analysis of large clinical corpora to test neurolinguistic hypotheses). Effective integration should target at least Modification level. From a constructivist perspective, LLM-based corpus analysis supports authentic learning through real clinical data, scaffolded learning via LLM-generated annotations students refine, and collaborative learning through group discussion of AI outputs. Communities of practice theory (Lave & Wenger, 1991) frames this as legitimate peripheral participation in professional linguistics communities.

3.3 Corpus Linguistics in Language-Related Education: Pedagogical Foundations and Innovations

Corpus linguistics has established significant pedagogical value across language-related disciplines, with data-driven learning (DDL) positioning students as researchers who discover linguistic patterns through direct corpus investigation rather than receiving pre-digested descriptions. This approach promotes inductive reasoning, pattern recognition, and empirical validation—cognitive skills essential for linguistic research and clinical practice (Boulton, 2012; Johns, 1991; 1998). Large language models enhance DDL by rapidly generating corpus patterns that students can examine, enabling focus on interpretation rather than data extraction.

Central to corpus-based education is the development of frequency and typicality awareness, as corpus analysis reveals which linguistic forms are common versus rare, typical versus marked. For clinical populations, corpus-based frequency norms inform assessment interpretation by determining whether a particular production pattern falls within or outside typical ranges for specific aphasia types (Berndt *et al*., 2000; Thompson et al., 2012). Students develop empirical grounding for clinical judgments through sustained corpus exposure.

Authentic language exposure distinguishes corpus-based education from constructed examples or idealized grammatical descriptions. Authentic clinical corpora reveal the complexity of real language production including disfluencies, self-corrections, and false starts, thereby preparing students for clinical realities (Gilquin & Granger, 2010). Exposure to variability within and across speakers develops realistic expectations and flexible analytical skills essential for professional practice. The investigative nature of corpus pedagogy encourages hands-on exploration where students actively query corpora, formulate hypotheses, test predictions, and refine analyses based on evidence. This investigative approach develops research literacy alongside linguistic knowledge, fostering skills transferable to professional contexts (Charles, 2015; Römer, 2011).

AI integration addresses many traditional barriers while creating new pedagogical possibilities. Reduced technical barriers emerge as natural language prompts replace complex query languages, making corpus analysis accessible to students without programming backgrounds. This accessibility enables accelerated analysis where students can examine multiple samples, compare across speakers or conditions, and iterate rapidly, transforming corpus analysis from isolated demonstration to integrated investigative method. Enhanced focus on interpretation occurs when annotation automation shifts student cognitive resources from mechanical coding to linguistic analysis and clinical reasoning. Students spend more time asking what patterns reveal about language processing rather than how to code particular constructions, aligning with constructivist principles emphasizing deep understanding over procedural execution. Integrated validation activities become pedagogically central as students compare LLM outputs against gold standards, identify systematic errors, and investigate why particular phenomena challenge automation. These metacognitive activities develop critical evaluation skills generalizable beyond corpus linguistics to any AI-enhanced professional context, while multilingual accessibility potentially expands through LLMs that handle diverse languages more flexibly than language-specific tools.

4. Design and Implementation Context

This paper presents a design-based framework developed and piloted across two Greek university programs during 2024–2025. The following sections describe the implementation context, corpus materials, and pedagogical design rationale. Full empirical evaluation of learning outcomes is ongoing and will be reported in subsequent work.

4.1 Research questions and aim

The overarching aim is to produce an evidence-based framework for LLM integration that balances technological innovation with pedagogical soundness while addressing discipline-specific needs.
RQ1: *To what extent can Large Language Models accurately perform multi-level linguistic annotation of Greek aphasic speech to support teaching scenarios and materials?*
RQ2: *Does integration of LLM-based corpus analysis enhance student learning outcomes compared to traditional corpus methods or no corpus instruction?*
RQ3: *What technical, pedagogical, and institutional barriers emerge during implementation, and what solutions prove effective?*
RQ4: *How do students perceive LLM-enhanced corpus analysis in terms of usability, learning value, and preparation for professional practice?*

4.2 Corpora Selection

The Greek CACLA (Corpus for Aphasic Clinical Language Analysis) serves as the primary corpus, comprising approximately 200 speech samples from Greek-speaking individuals with aphasia representing diverse types and severity levels (Papathanasiou *et al*., 2017). We curated a pedagogical subset of 36 transcripts (12 for demonstrations, 24 for student projects) selected for: (1) appropriate length (150-400 words); (2) linguistic representativeness of different aphasia types; (3) transcription quality following CHAT conventions; (4) task diversity; and (5) complexity gradient enabling differentiated instruction.

Gold standard annotations were developed through multi-stage expert consensus, with two linguists independently annotating each transcript following Universal Dependencies guidelines for Greek (Prokopidis & Papageorgiou, 2017) and clinical error coding systems (Schwartz *et al*., 2009). Disagreements (approximately 12%) were resolved through adjudication.

To accommodate different pedagogical objectives across programs, we incorporated comparison datasets enabling contrastive analysis. For English/Linguistics students, we included a parallel corpus of 10 transcripts from neurologically intact Greek speakers performing identical elicitation tasks (picture description, narrative retelling, procedural discourse). This typical language corpus enables students to systematically compare linguistic patterns between aphasic and normal speech—identifying which features represent pathological deviation versus normal variability, examining quantitative differences in productivity and complexity measures, and developing empirically-grounded understanding of linguistic breakdown patterns (Armstrong, 2000; Goodglass *et al*., 2001). This contrastive approach aligns with computational linguistics pedagogy emphasizing comparative analysis and data-driven pattern recognition (Römer, 2011). Additionally, we incorporated AphasiaBank English samples for cross-linguistic comparison (MacWhinney et al., 2011), and outputs from traditional NLP tools (Greek spaCy, Stanza) for comparative evaluation of LLM versus rule-based/statistical annotation approaches—particularly relevant for students focusing on computational methods.

4.3 Study design behind teaching procedures

The study employs a quasi-experimental pre-post design with comparison groups implemented across two Greek university departments during 2024-2025. The first group (n ≈ 55) receives a full LLM-enhanced curriculum including prompt engineering, validation methodology, and collaborative analysis projects. The second group (n ≈ 67) receives AI-modified clinical training with systematic corpus analysis and aphasic speech data in two different modules.

All groups complete identical pre/post tests assessing corpus linguistics knowledge, clinical language understanding, and analytical skills. Data collection employs convergent mixed-methods design (Creswell & Plano Clark, 2017): quantitative measures (knowledge tests, annotation accuracy, Likert-scale surveys) and qualitative data (focus groups, observation checklists). Statistical analysis includes mixed-model ANOVAs examining pre-post gains across groups, t-tests comparing practical competency, and regression models predicting outcomes from background variables. Qualitative data undergoes thematic analysis following Braun and Clarke (2006).

4.4 Participants

Participants comprise 122 third- and fourth-year undergraduate students across two programs: School of English (n = 55) and Speech-Language Pathology (n = 67). The sample is predominantly female (85%), aged 20-26 years (M = 21.8), consistent with linguistics and SLP program demographics internationally. All are native Greek speakers or have advanced proficiency (C1/C2; 9 bilinguals with Russian, Albanian, Ukrainian and Turkish). Prior coursework includes foundational linguistics, though only 23% report corpus linguistics exposure and 12% have programming experience. Technology experience is mixed: 67% have used ChatGPT (or any other LLM) for general purposes, but only 15% for academic analytical tasks. This profile suggests students can navigate AI interfaces but lack critical evaluation skills for linguistic applications—a key learning objective.

5. Educational Applications

5.1 Curriculum Integration

The LLM-based corpus analysis framework integrates across three distinct but complementary curricular modules, each addressing specific disciplinary competencies while leveraging shared technological infrastructure. The Theoretical Module (*Aphasia and Related Disorders*) emphasizes neurocognitive models of language breakdown, classification and symptomatology of acquired aphasia and related disorders, and evidence-based principles of language assessment and rehabilitation—developing foundational theoretical competence in adult neurogenic communication disorders. The Digital Module (*Computational Linguistics and Natural Language Processing*) emphasizes annotation methodology, prompt engineering for linguistic tasks, and systematic evaluation of automated analysis tools—developing technical competencies in NLP applications. The Clinical Module (*Clinical Practice III*) emphasizes supervised application of assessment, differential diagnosis, intervention planning, and professional documentation in real clinical settings—developing advanced clinical reasoning, ethical decision-making, and entry-level professional competence in adult speech-language pathology. This module enables students to consolidate theoretical knowledge and digital competencies into advanced, context-sensitive clinical decision-making. This three-pronged approach ensures students develop theoretical understanding, technical proficiency, and clinical reasoning—essential competencies for contemporary speech-language pathology and computational linguistics professionals (ASHA, 2016; Jurafsky & Martin, 2023).

5.2 Specific Educational Modules

5.2.1 Theoretical Module: Aphasia and Related Disorders

The Theoretical Module Aphasia and Related Disorders provides the conceptual and neurocognitive foundation for the integrated curriculum and is implemented as a structured lecture sequence (9 instructional hours; for more details see Appendix A) designed to develop theoretical coherence, analytic depth, and preparedness for advanced clinical application. The module addresses a longstanding challenge in speech-language pathology education: enabling students to move beyond surface-level classification toward principled understanding of how neurological damage disrupts language systems across modalities and levels of linguistic organization (Goodglass & Wingfield, 1997; Caplan & Marshall,1992). In Session 1, Foundations of Aphasia and Neurogenic Language Breakdown, students are introduced to historical and contemporary definitions of aphasia, etiological factors, and neuroanatomical correlates of language processing. Classical aphasia syndromes are presented alongside modern neurocognitive and functional models, highlighting both their heuristic value and their limitations in capturing real-world communicative functioning (Bastiaanse & Prins, 2013; Dronkers et al., 2017). Through analysis of authentic aphasic language samples, students begin to recognize inter- and intra-speaker variability and to critically evaluate the adequacy of syndromic labels. This session supports a shift from taxonomic memorization to explanatory, model-based reasoning about language breakdown. Session 2, Language Assessment and Breakdown Patterns, focuses on systematic evaluation of spoken and written language across comprehension, expression, naming, repetition, reading, writing, and discourse. Students examine standardized and non-standardized assessment approaches and learn to interpret error patterns in relation to underlying linguistic and cognitive processes, such as lexical access, morphosyntactic encoding, and discourse planning (Chapey, 2001). Particular emphasis is placed on distinguishing aphasia from co-occurring motor speech and cognitive-communication disorders, reinforcing diagnostic precision and reducing misclassification in clinical contexts. In Session 3, Aphasia, Related Disorders, and Functional Communication, the focus expands to acquired language disorders associated with traumatic brain injury and dementia. Students explore how diffuse versus focal neuropathology affects language, cognition, and communicative participation, and how progressive versus non-progressive conditions shape assessment and intervention priorities (Chapey, 2001; Bayles & Tomoeda, 2014). The session cites aphasia within the WHO–ICF framework, emphasizing activity limitations and participation restrictions alongside impairment-level analysis (WHO, 2001). Conceptual principles of intervention are introduced, highlighting evidence-based practice, functional goal setting, and the role of communication partners (Simmons-Mackie et al., 2010). Pedagogically, this module prioritizes conceptual integration and theoretical grounding over procedural skill acquisition. By developing explanatory models of language breakdown, it provides the cognitive scaffolding required for both computational analysis in the Digital Module and applied clinical reasoning in Clinical Practice III, supporting deep learning and transfer across contexts (National Research Council,, 2000).

5.2.2 Digital Module: Computational Linguistics and Natural Language Processing

The Digital Module represents the core computational linguistics application of LLM-based corpus analysis, implemented as a four-session sequence (5 instructional hours; for more details see Appendix A) designed to develop both technical proficiency and critical evaluation skills. This module addresses a critical gap in linguistics education: making sophisticated corpus annotation accessible to students without programming backgrounds while maintaining methodological straightforwardness. In session 1, Foundations and Initial Exploration introduces students to the CACLA corpus structure, Greek morphosyntactic complexity, and traditional annotation challenges. Students conduct manual annotation of a brief sample (15-20 words), documenting time investment and difficulties

encountered. This experiential foundation establishes appreciation for automation benefits while grounding students in linguistic fundamentals—preventing what Pea (2004) terms "cognitive offloading" where technology bypasses essential conceptual development. The session concludes with a live demonstration of Claude performing identical annotation tasks in seconds, creating productive cognitive dissonance that motivates deeper investigation of how LLMs accomplish linguistic analysis.

During session 2, Prompt Engineering and Annotation Generation shifts to hands-on LLM interaction. Working in small groups, students design prompts for morphological annotation (POS tagging, lemmatization, feature extraction) using provided templates as starting points. The iterative prompt refinement process—testing, evaluating output quality, modifying instructions, retesting—develops metacognitive awareness about linguistic specification and computational interpretation (White *et al*., 2023). Students discover that effective prompts require explicit articulation of annotation conventions, output formats, and edge case handling—essentially requiring them to formalize their linguistic knowledge computationally. This "teaching the machine" process deepens linguistic understanding while developing prompt engineering as an emerging professional skill (Kasneci *et al*., 2023).

In the 3rd session, Validation and Critical Evaluation represents the pedagogically crucial component distinguishing our approach from uncritical AI adoption. Students systematically compare their LLM-generated annotations against expert gold standards, calculating precision, recall, and F1-scores for different annotation types. Error analysis reveals systematic patterns: LLMs excel at standard morphological forms but struggle with neologisms, phonemic paraphasias, and agrammatic structures lacking function words (Kuzman *et al*., 2023). Students categorize errors, hypothesize about computational causes (training data bias, context limitations, lack of clinical linguistic knowledge), and discuss implications for research applications. This critical engagement prevents "automation bias"—the tendency to over-trust automated outputs—while developing an understanding of AI capabilities and constraints (Goddard *et al*., 2012).

Finally, during session 4, Comparative Analysis and Synthesis culminates in contrastive examination of aphasic versus typical language corpora. Students use their validated LLM workflows to analyze multiple samples, extracting quantitative measures (MLU, TTR, grammatical accuracy rates) and identifying qualitative patterns. The larger dataset sizes enabled by automation allow observation of cross-speaker generalizations impossible with manual methods within course timeframes. Final presentations synthesize technical findings with linguistic interpretation: What do corpus patterns reveal about Greek morphological processing? How do computational analysis results align with or challenge theoretical models of agrammatism? This integration of technical and theoretical dimensions exemplifies Bloom's highest taxonomic levels—synthesis and evaluation (Anderson & Krathwohl, 2001). Pedagogical significance lies in balancing automation efficiency with analytical depth, technical skill development with critical thinking, and computational methods with linguistic insight—preparing students for futures where AI tools are ubiquitous but human expertise remains essential.

5.2.3 Clinical Module: Clinical Practice III

The Clinical Module *Clinical Practice III* represents the culminating application of theoretical and analytical competencies within supervised real-world clinical environments. Implemented as a semester-long clinical placement with supporting lectures (10 ECTS), the module is designed to support students' transition from guided learning to entry-level professional practice. Its central pedagogical objective is the development of integrated clinical reasoning, ethical judgment, and professional autonomy under supervision, consistent with international standards for clinical education in speech-language pathology (ASHA, 2016; RCSLT, 2017). In Session 1, Advanced Clinical

Assessment and Case Formulation, students engage in comprehensive evaluation of adults with acquired communication disorders, including structured case history intake, orofacial and cranial nerve examination, cognitive screening, and detailed language assessment. Drawing on theoretical knowledge from the *Aphasia and Related Disorders* module, students synthesize linguistic, cognitive, and neurological data into coherent clinical hypotheses. Emphasis is placed on differential diagnosis in complex cases involving overlapping aphasic, cognitive-communication, and motor speech features, reflecting best practices in adult neurogenic assessment (Papathanasiou, Coppens, & Potagas, 2017). Session 2, Clinical Decision-Making and Intervention Planning, focuses on translating assessment findings into evidence-based therapy goals and intervention plans. Students are trained to prioritize functional communication outcomes, align goals with patient needs and contextual factors, and justify clinical decisions using empirical evidence and theoretical rationale (Chapey, 2001). Supporting lectures on acquired language disorders in traumatic brain injury and dementia reinforce the importance of flexible clinical reasoning and longitudinal planning in progressive and non-progressive conditions. In Session 3, Documentation, Outcome Monitoring, and Professional Practice, students develop advanced skills in clinical documentation, including evaluation reports, therapy plans, progress notes, and SOAP documentation. Cognitive and quality-of-life questionnaires are incorporated into outcome monitoring, reinforcing a holistic and patient-centered approach to clinical effectiveness (Hilari & Byng, 2009). Ethical considerations—including informed consent, confidentiality, professional accountability, and interdisciplinary collaboration—are embedded throughout clinical practice, aligning with professional codes of ethics and legal frameworks (ASHA, 2016). Pedagogically, Clinical Practice III exemplifies experiential learning at the highest level, integrating theoretical knowledge, analytical competence, and procedural skill within authentic clinical contexts. By the end of the module, students demonstrate readiness for professional practice, characterized by evidence-based clinical reasoning, ethical awareness, reflective capacity, and adaptability to complex real-world communication disorders.

5.3 Towards a common educational scenario template

The proposed educational scenario template provides a flexible, adaptable framework for integrating LLMs and corpus analysis across diverse linguistic and clinical contexts beyond the specific CACLA implementation. This modular template follows a four-phase pedagogical progression that can be customized for different languages, corpora, and target competencies: Phase 1 (Foundation and Manual/Traditional Exploration) establishes baseline understanding through hands-on engagement with the chosen corpus and traditional annotation methods; Phase 2 (AI-Enhanced Pipeline Development) introduces LLM integration through systematic prompt engineering and automation workflows; Phase 3 (Validation and Critical Analysis) develops evaluation competencies through systematic comparison with gold standards and error analysis; and Phase 4 (Synthesis and Application) culminates in comprehensive analysis and presentation of findings. The template's strength lies in its adaptability—instructors can substitute alternative corpora (clinical, literary, social media, multilingual), modify linguistic annotation levels (phonological, morphological, syntactic, discourse), adjust technical complexity for different student populations, and customize assessment criteria for specific learning objectives. Essential template components include prerequisite specification, resource requirements, step-by-step process documentation, variation suggestions for different contexts, and assessment rubrics, enabling educators to create robust LLM-enhanced corpus linguistics curricula tailored to their institutional needs and student populations while maintaining pedagogical rigor and critical thinking development.

6. Discussion

6.1 Pedagogical Rationale and Anticipated Outcomes

The Theoretical Module was designed to foster foundational competencies that transformed their understanding of neurological language disorders from superficial classification to principled theoretical reasoning. Through systematic engagement with Session 1 (Foundations of Aphasia and Neurogenic Language Breakdown), students will learn to critically evaluate classical aphasia syndromes alongside modern neurocognitive models, recognizing both their heuristic value and limitations in capturing real-world communicative functioning. The analysis of authentic aphasic language samples will enable students to appreciate inter- and intra-speaker variability while moving beyond taxonomic memorization toward explanatory, model-based reasoning about language breakdown. Session 2 (Language Assessment and Breakdown Patterns) will equip students with systematic evaluation skills across comprehension, expression, and discourse modalities, teaching them to interpret error patterns in relation to underlying linguistic and cognitive processes such as lexical access and morphosyntactic encoding. Most significantly, Session 3 (Aphasia, Related Disorders, and Functional Communication) will broaden students' clinical perspective to encompass acquired language disorders in traumatic brain injury and dementia, while situating aphasia within the WHO-ICF framework to emphasize activity limitations and participation restrictions alongside impairment-level analysis.

The Digital Module will successfully bridge the gap between theoretical linguistic knowledge and practical computational application, enabling students to develop both technical proficiency and critical evaluation expertise. Through Session 1 (Foundations and Initial Exploration), students will gain experiential understanding of annotation challenges by manually processing brief samples before witnessing LLM capabilities, creating productive cognitive dissonance that will motivate deeper investigation of computational linguistic analysis. Session 2 (Prompt Engineering and Annotation Generation) will teach students to articulate their linguistic knowledge computationally through iterative prompt refinement, discovering that effective prompts require explicit specification of annotation conventions, output formats, and edge case handling—essentially formalizing their linguistic understanding for computational interpretation. The pedagogically crucial Session 3 (Validation and Critical Evaluation) will develop systematic comparison skills as students calculated precision, recall, and F1-scores while conducting error analysis that will reveal LLM strengths in standard morphological forms and limitations with neologisms, phonemic paraphasias, and agrammatic structures. Session 4 (Comparative Analysis and Synthesis) will enable students to leverage automation for examining larger datasets, extracting quantitative measures like MLU and TTR while synthesizing technical findings with linguistic interpretation to address questions about Greek morphological processing and theoretical models of agrammatism.

Students in Clinical Practice III will achieve the critical transition from guided academic learning to entry-level professional practice through integration of theoretical knowledge, analytical competencies, and supervised clinical experience. Session 1 (Advanced Clinical Assessment and Case Formulation) will teach students to synthesize linguistic, cognitive, and neurological data into coherent clinical hypotheses while engaging in comprehensive evaluation including case history intake, orofacial examination, and detailed language assessment. The emphasis on differential diagnosis in complex cases involving overlapping aphasic, cognitive-communication, and motor speech features will prepare students for real-world clinical complexity. Session 2 (Clinical Decision-Making and Intervention Planning) developed evidence-based clinical reasoning as students will learn to translate assessment findings into functional therapy goals, prioritize communication outcomes, and justify clinical decisions using empirical evidence and theoretical rationale. Session 3 (Documentation, Outcome Monitoring, and Professional Practice) will equip students with advanced clinical documentation skills including evaluation reports and SOAP documentation, while embedding ethical considerations throughout clinical practice. By module completion, students will demonstrate

readiness for professional practice characterized by evidence-based clinical reasoning, ethical awareness, reflective capacity, and adaptability to complex real-world communication disorders.

6.2 Educational Effectiveness

The advantages outlined in section 6.3 demonstrate LLMs' transformative potential for corpus-based education in linguistics and speech-language pathology. Primary benefits include democratized accessibility through elimination of programming prerequisites and natural language prompting that replaces complex query languages, enabling sophisticated corpus analysis across diverse technical backgrounds. Time efficiency gains facilitate qualitative pedagogical shifts from analyzing 2-3 brief samples to examining dozens of transcripts, revealing cross-speaker patterns essential for clinical competence while redirecting cognitive resources from mechanical coding toward higher-order linguistic analysis and clinical reasoning. For clinical applications, LLMs provide rapid extraction of quantitative language measures that inform evidence-based assessment and intervention outcomes, with particular value for under-resourced languages like Greek where specialized NLP tools remain limited. The pedagogical flexibility allows instructors to customize difficulty and scaffold complexity progressively, making LLM integration essential for interdisciplinary contexts spanning computational linguistics to clinical practice.

6.3 LLMs for Corpus Analysis, Language and Speech Therapy: Advantages

Large Language Models offer transformative advantages for corpus-based education in linguistics and speech-language pathology that address longstanding pedagogical barriers. Accessibility and democratization represent primary benefits: LLMs eliminate programming prerequisites, making sophisticated corpus analysis available to students across diverse technical backgrounds (Kasneci et al., 2023). Natural language prompting replaces complex query languages, lowering entry barriers that historically limited corpus linguistics to computationally-trained specialists (Reppen, 2010).

Time efficiency gains enable qualitative pedagogical shifts. Where manual annotation constrained students to analyzing 2-3 brief samples, LLM-assisted workflows permit examination of dozens of transcripts, revealing cross-speaker patterns and population-level generalizations essential for clinical and research competence (Boulton & Cobb, 2017). This scalability transforms corpus analysis from isolated demonstration to integrated investigative method. Enhanced focus on interpretation emerges as automation handles mechanical coding, redirecting cognitive resources toward linguistic analysis, clinical reasoning, and theoretical synthesis—higher-order thinking central to professional education (Anderson & Krathwohl, 2001; Fink, 2013).

For clinical applications, LLMs provide rapid extraction of quantitative language measures (productivity, complexity, accuracy indices) that inform evidence-based assessment, benchmark intervention outcomes, and support clinical documentation requirements increasingly demanded in healthcare contexts (ASHA, 2016; Thompson *et al*., 2012). Multilingual capabilities particularly benefit under-resourced languages like Greek, where specialized NLP tools remain limited; multilingual LLMs offer reasonable performance without extensive language-specific development (Joshi et al., 2020). Finally, pedagogical flexibility allows instructors to customize difficulty, scaffold complexity progressively, and adapt activities to diverse learning objectives—essential for interdisciplinary contexts spanning computational linguistics to clinical practice.

6.4 Limitations and Challenges

Despite significant advantages, LLM integration presents substantial limitations requiring careful pedagogical management. Technical limitations constrain reliability for critical applications. Annotation accuracy varies considerably by linguistic level and phenomenon: while morphological tagging achieves 85-92% accuracy for standard Greek forms, performance degrades substantially for clinical language features including neologisms (62% accuracy), phonemic paraphasias (71%), and agrammatic structures with systematic omissions (68%) (cf. Kuzman et al., 2023; Liu et al., 2024). Context understanding remains shallow; LLMs lack genuine comprehension of clinical presentations, neuroanatomical substrates, or theoretical frameworks guiding interpretation. Hallucination risks—confident generation of plausible but incorrect analyses—pose serious concerns when students lack expertise to recognize errors (Alkaissi & McFarlane, 2023).

Educational considerations demand vigilant attention. Over-reliance on automation without developing foundational skills represents the primary pedagogical risk; students may generate sophisticated-appearing analyses without grasping underlying linguistic principles—what Salomon et al. (1991) term "intelligence in the system rather than intelligence with the system." Critical thinking development requires deliberate scaffolding; validation exercises, error analysis, and comparative evaluation must be integrated systematically rather than optional supplements. Technical literacy requirements extend beyond operational skills to conceptual understanding of LLM capabilities, limitations, and appropriate use cases—a new form of literacy not yet systematically addressed in linguistics curricula (Cardon *et al*., 2023).

Practical barriers complicate implementation. Cost considerations vary: while free LLM tiers enable basic exploration, intensive educational use may require institutional subscriptions ($20-30/student/month), substantial in resource-constrained contexts. Privacy and data security concerns arise with clinical corpora containing sensitive information, even when de-identified; institutional policies may restrict uploading patient-derived data to commercial AI platforms. Instructor expertise gaps emerge as most linguistics faculty lack experience with prompt engineering, AI evaluation methodology, or pedagogical strategies for critical technology engagement—necessitating substantial professional development investment (Mishra & Koehler, 2006; Zawacki-Richter et al., 2019).

6.5 Best Practices for Implementation

Evidence from our implementation suggests several critical practices for successful LLM integration. Foundational skills first: Students must develop basic linguistic analysis competencies through manual annotation before automation introduction, ensuring conceptual understanding precedes technological efficiency (Pea, 2004). Explicit prompt engineering requries students to record, version, and justify all prompts develops metacognitive awareness about linguistic specification and creates reproducible workflows aligned with research integrity standards (White et al., 2023). Scaffolded complexity progression is eminent by beginning with straightforward annotation tasks (POS tagging) before advancing to challenging phenomena (discourse coherence, clinical error detection) builds confidence while revealing AI limitations organically. Finally, interdisciplinary collaboration will rise by pairing computational linguistics and clinical students leverages complementary expertise, models professional teamwork, and enriches learning through diverse perspectives.

6.6 Future Directions

Future development should address several critical directions. Longitudinal evaluation tracking whether corpus analysis skills and critical AI literacy persist beyond course completion and transfer to professional practice remains essential for validating educational impact (Kirkpatrick & Kirkpatrick, 2006). Cross-linguistic expansion to additional under-resourced languages could democratize corpus linguistics globally, though requiring systematic validation of LLM performance across diverse

linguistic typologies (Joshi *et al*., 2020). Hybrid approaches integrating LLMs with specialized clinical NLP tools may optimize accuracy-efficiency trade-offs by using LLMs for initial annotation and traditional tools for validation, or vice versa (Bryant *et al*., 2023). Custom fine-tuning of open-source models on clinical language corpora could enhance performance on atypical language while addressing privacy concerns through local deployment, though requiring institutional computational infrastructure (Üstün *et al*., 2024).

To achieve curriculum standardization, it is needed to develop consensus guidelines for AI literacy competencies in linguistics and SLP programs would ensure systematic preparation for technology-integrated professional futures. Ethical framework development specifically addressing AI use with vulnerable populations, clinical data, and healthcare applications remains urgent as technology outpaces regulatory guidance (Mittelstadt *et al*., 2016). Finally, authentic clinical integration can be achieved by moving beyond educational simulation to implementing LLM-assisted corpus analysis in actual clinical assessment, treatment planning, and outcome documentation would demonstrate real-world value while identifying practical implementation barriers. Collaborative research between educational institutions, healthcare facilities, and AI developers could accelerate responsible clinical translation benefiting both professional training and patient care.

7. Conclusions

The three-module structure — Theoretical, Computational, and Clinical — provides a principled progression from foundational neurolinguistic knowledge through computational methodology to supervised clinical application, while the four-phase scenario template gives instructors in diverse institutional contexts a reusable structure adaptable to other languages, corpora, and learning objectives. Throughout, the framework positions LLMs as "digital co-educators" rather than authoritative tools: by requiring students to validate automated outputs against gold standards and to analyse systematic errors, it builds critical evaluation skills and technological literacy alongside disciplinary knowledge, counteracting the automation bias that uncritical AI adoption risks.

Several limitations should be acknowledged; full empirical evaluation of learning outcomes is ongoing, and the current implementation is confined to two Greek university programs; generalisability to other national and linguistic contexts remains to be established. Reliance on commercial LLM platforms also raises sustainability and data-privacy concerns when working with patient-derived clinical material, and successful implementation presupposes instructors with competence in both prompt engineering and clinical linguistics — expertise not yet widely available in linguistics faculties.

For educators, the framework offers a concrete, theoretically grounded entry point for curriculum development that requires no programming background in instructors or students. For the field more broadly, it models responsible AI integration in clinical professional education — treating technological literacy as inseparable from disciplinary expertise and ethical awareness. Priorities for future work include longitudinal evaluation of skill transfer to professional practice, cross-linguistic validation, and the development of open-source or locally deployable alternatives to commercial LLMs for use with sensitive clinical corpora.

Acknowledgments

The research project was supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the "Basic Research Financing (Horizontal Support for All Sciences), National Recovery and Resilience Plan (Greece 2.0)" (Project Number: 016344). The principal investigator was Ilias Papathanasiou.

References

American Speech-Language-Hearing Association [ASHA]. (2016). *Scope of practice in speech-language pathology*. Retrieved from https://www.asha.org/policy/

American Speech-Language-Hearing Association [ASHA]. (2023). *2023 membership and affiliation profile*. https://www.asha.org/

Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman.

Anil, R., Dai, A. M., Firat, O., … & Wu, Y. (2023). PaLM 2 technical report. *arXiv preprint*. https://doi.org/10.48550/arXiv.2305.10403

Anthropic. (2024). *Claude 3 model card*. Retrieved from https://www.anthropic.com

Armstrong, E. (2000). Aphasic discourse analysis: The story so far. *Aphasiology*, 14(9), 875-892.

Ball, M. J. (2013). *Research methods in clinical linguistics and phonetics: A practical guide*. Wiley-Blackwell.

Ball, M. J., Perkins, M. R., Müller, N., & Howard, S. (Eds.). (2008). *The Handbook of Clinical Linguistics*. Blackwell Publishing.

Barbieri, F., & Eckhardt, S. (2007). Applying corpus-based findings to form-focused instruction: The case of reported speech. *Language Teaching Research*, 11(3), 319-346. https://doi.org/10.1177/136216880707756

Bardovi-Harlig, K., & Mossman, S. (2016). Corpus-based materials development for teaching and learning pragmatic routines. In B. Tomlinson (Ed.), *SLA research and materials development for language learning* (pp. 250-267). Routledge.

Bastiaanse, R., & Edwards, S. (2004). Word order and finiteness in Dutch and English Broca's and Wernicke's aphasia. *Brain and Language*, 89(1), 91-107. https://doi.org/10.1016/S0093-934X(03)00306-7

Bastiaanse, R., & van Zonneveld, R. (2005). Sentence production with verbs of alternating transitivity in agrammatic Broca's aphasia. *Journal of Neurolinguistics*, 18(1), 57-66. https://doi.org/10.1016/j.jneuroling.2004.11.006

Bastiaanse, R., & Prins, R. S. (2013). Aphasia. In L. Cummings (Ed.), *The Cambridge handbook of communication disorders* (pp. 224–246). Cambridge University Press.

Bayles, K. A., & Tomoeda, C. K. (2014). *Cognitive-communication disorders of dementia: Definition, diagnosis, and treatment* (2nd ed.). Plural Publishing Inc.

Beam, A. L., & Kohane, I. S. (2018). Big data and machine learning in health care. *JAMA*, 319(13), 1317-1318. https://doi.org/10.1001/jama.2017.18391

Belland, B. R., Kim, C., & Hannafin, M. J. (2013). A framework for designing scaffolds that improve motivation and cognition. *Educational Psychologist*, 48(4), 243-270. https://doi.org/10.1080/00461520.2013.838920

Berndt, R. S., Wayland, S., Rochon, E., Saffran, E., & Schwartz, M. (2000). *Quantitative production analysis*. Psychology Press.

Bhat, S., Culotta, A., & Bhamidipati, N. (2022). *DisfluencyFixer: A tool to enhance Language Learning through Speech To Speech Disfluency Correction*. arXiv preprint. arXiv:2305.16957. https://doi.org/10.48550/arXiv.2305.16957

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.

Bitzenbauer, P. (2023). ChatGPT in physics education: A pilot study on easy-to-implement activities. *Contemporary Educational Technology*, 15(3), ep430. https://doi.org/10.30935/cedtech/13176

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook 1: Cognitive domain*. David McKay.

Boersma, P., & Weenink, D. (2021). *Praat: Doing phonetics by computer* [Computer program]. Version 6.4.60.

Bommasani, R., Hudson, D. A., Adeli, E., et al. & Liang, P. (2021). *On the opportunities and risks of foundation models*. arXiv preprint arXiv:2108.07258. https://doi.org/10.48550/arXiv.2108.07258

Boulton, A. (2012). Corpus consultation for ESP: A review of empirical research. In A. Boulton, S. Carter-Thomas, & E. Rowley-Jolivet (Eds.), *Corpus-informed research and learning in ESP: Issues and applications* (pp. 261-291). John Benjamins. https://doi.org/10.1075/scl.52.11bou

Boulton, A., & Cobb, T. (2017). Corpus use in language learning: A meta-analysis. *Language Learning*, 67(2), 348-393. https://doi.org/10.1111/lang.12224

Brackenbury, T., Burroughs, E., & Hewitt, L. E. (2008). A qualitative examination of current guidelines for evidence-based practice in child language intervention. *Language, Speech, and Hearing Services in Schools*, 39(1), 78-88. https://doi.org/10.1044/0161-1461(2008/008)

Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds.). (2000). *How people learn: Brain, mind, experience, and school* (Expanded ed.). National Academy Press.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77-101. https://doi.org/10.1191/1478088706qp063oa

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., …D. Amodei. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877-1901. https://doi.org/10.48550/arXiv.2005.14165

Bryant, C., Yuan, Z., Qorib, M. R., et al. (2023). Grammatical error correction: A survey of the state of the art. *Computational Linguistics*, 49(3), 1-59. https://doi.org/10.1162/coli_a_00478

Bubeck, S., Chandrasekaran, V., Eldan, R., …, Zhang, L. (2023). *Sparks of artificial general intelligence: Early experiments with GPT-4*. arXiv preprint arXiv:2303.12712. https://doi.org/10.48550/arXiv.2303.12712

Caplan, D. (1987). *Neurolinguistics and linguistic aphasiology: An introduction*. Cambridge University Press.

Caplan, D., & Marshall, J. C. (Eds.) (1992). *Language: Structure, processing, and disorders*. The MIT Press.

Cardon, P. W., Fleischmann, C., Aritz, J., Logemann, M., & Heidewald, J. (2023). The challenges and opportunities of AI-assisted writing: Developing AI literacy for the AI age. *Business and Professional Communication Quarterly*, 86(3), 257-295. https://doi.org/10.1177/23294906231176

Chan, C. K. Y., & Hu, W. (2023). Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education*, 20(1), 43. https://doi.org/10.1186/s41239-023-00411-8

Chapey, R. (2001). *Language intervention strategies in aphasia and related neurogenic communication disorders* (4th ed.). Lippincott Williams & Wilkins.

Charles, M. (2015). Same task, different corpus: The role of personal corpora in EAP classes. In A. Leńko-Szymańska & A. Boulton (Eds.), *Multiple affordances of language corpora for data-driven learning* (pp. 131-154). John Benjamins. https://doi.org/10.1075/scl.69.07cha

Covington, M. A., & McFall, J. D. (2010). Cutting the Gordian knot: The moving-average type-token ratio (MATTR). *Journal of Quantitative Linguistics*, 17(2), 94-100. https://doi.org/10.1080/09296171003643098

Creswell, J. W., & Plano Clark, V. L. (2017). *Designing and conducting mixed methods research* (3rd ed.). SAGE.

Crosthwaite, P. (2021). *Data-driven learning for the next generation: Corpora and DDL for pre-tertiary learners*. Routledge.

Crystal, D. (2013). *Profiling linguistic disability* (3rd ed.). Wiley-Blackwell.

Dalton, S. G., & Richardson, J. D. (2019). A large-scale comparison of main concept production between persons with aphasia and persons without brain injury. *American Journal of Speech-Language Pathology*, 28(1S), 293-320. https://doi.org/10.1044/2018_AJSLP-17-0166

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT* (pp. 4171-4186). https://doi.org/10.18653/v1/N19-1423

Dillenbourg, P. (1999). What do you mean by collaborative learning? In P. Dillenbourg (Ed.*), Collaborative-learning: Cognitive and computational approaches* (pp. 1-19). Elsevier.

Dillenbourg, P., Järvelä, S., & Fischer, F. (2016). *The evolution of research on computer-supported collaborative learning*. Springer.

Dronkers, N. F., Ivanova, M. V., & Baldo, J. V. (2017). What do language disorders reveal about brain–language relationships? From classic models to network approaches. *Journal of the International Neuropsychological Society*, *23*(9–10), 741–754. https://doi.org/10.1017/S1355617717001126

Ettinger, A. (2020). What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8, 34-48. https://doi.org/10.48550/arXiv.1907.13528

Fergadiotis, G., & Wright, H. H. (2011). Lexical diversity for adults with and without aphasia across discourse elicitation tasks. Aphasiology, 25(11), 1414-1430. https://doi.org/10.1080/02687038.2011.603898

Fergadiotis, G., Wright, H. H., & Capilouto, G. J. (2011). Productive vocabulary across discourse types. *Aphasiology*, 25(10), 1261-1278. https://doi.org/10.1080/02687038.2011.606974

Fink, L. D. (2013). *Creating significant learning experiences: An integrated approach to designing college courses* (2nd ed.). Jossey-Bass.

Friginal, E., & Hardy, J. A. (Eds.). (2014). *Corpus-based sociolinguistics: A guide for students*. Routledge.

Gilquin, G., & Granger, S. (2010). How can data-driven learning be used in language teaching? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 359-370). Routledge.

Goodglass, H. (1993). *Understanding aphasia*. Academic Press.

Goodglass, H., & Wingfield, A. (Eds.). (1997). *Anomia: Neuroanatomical and cognitive correlates*. Academic Press.

Goodglass, H., Kaplan, E., & Barresi, B. (2001). *Boston diagnostic aphasia examination* (3rd ed.). Lippincott Williams & Wilkins.

Grogan-Johnson, S., Gabel, R. M., Taylor, J., Rowan, L. E., Alvares, R., & Schenker, J. (2013). A pilot investigation of speech sound disorder intervention delivered by telehealth to school-age children. *International Journal of Telerehabilitation*, 3(1), 31-42. https://doi.org/10.5195/ijt.2011.6064

Hamilton, E. R., Rosenberg, J. M., & Akcaoglu, M. (2016). The substitution augmentation modification redefinition (SAMR) model: A critical review and suggestions for its use. *TechTrends*, 60(5), 433-441. https://doi.org/10.1007/s11528-016-0091-y

Hao, Y., Mendelsohn, J., Sterneck, R., Martinez, R., & Frank, R. (2021). Probabilistic predictions of people perusing: Evaluating metrics of language model performance for psycholinguistic modelling. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics* (pp. 75-86). https://doi.org/10.18653/v1/2020.cmcl-1.10

Harris, J., Mishra, P., & Koehler, M. (2009). Teachers' technological pedagogical content knowledge and learning activity types: Curriculum-based technology integration reframed. *Journal of Research on Technology in Education*, 41(4), 393-416. https://doi.org/10.1080/15391523.2009.10782536

Herring, M. C., Koehler, M. J., & Mishra, P. (Eds.). (2016). *Handbook of technological pedagogical content knowledge (TPACK) for educators* (2nd ed.). Routledge.

Herrington, J., & Oliver, R. (2000). An instructional design framework for authentic learning environments. *Educational Technology Research and Development*, 48(3), 23-48. https://doi.org/10.1007/BF02319856

Hilari, K., & Byng, S. (2009). Health-related quality of life in people with severe aphasia. *International Journal of Language & Communication Disorders*, *44*(2), 193–205. https://doi.org/10.1080/13682820802008820

Hill, A. J., & Theodoros, D. (2017). Research into telehealth applications in speech-language pathology. *Journal of Telemedicine and Telecare*, 23(1), 37-47. https://doi.org/10.1258/135763302320272158

Hill, A. J., Theodoros, D., Russell, T., & Ward, E. (2020). Using telerehabilitation to deliver evidence-based speech-language pathology services. In R. Swisher & L. T. Goldberg (Eds.), *Innovations in allied health fieldwork education* (pp. 145-162). Springer.

Hillis, A. E. (2007). Aphasia: Progress in the last quarter of a century. *Neurology*, 69(2), 200-213. https://doi.org/10.1212/01.wnl.0000265600.69385.6f

Hilton, J. T. (2016). A case study of the application of SAMR and TPACK for reflection on technology integration into two social studies classrooms. *Social Studies*, 107(2), 68-73. https://doi.org/10.1080/00377996.2015.1124376

Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial intelligence in education: Promises and implications for teaching and learning*. Center for Curriculum Redesign.

Holton, D., Mackridge, P., & Philippaki-Warburton, I. (2004). *Greek: A comprehensive grammar of the modern language*. Routledge.

Honnibal, M., & Montani, I. (2017). *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*. https://spacy.io

Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H., & Aerts, H. J. (2018). Artificial intelligence in radiology. *Nature Reviews Cancer*, 18(8), 500-510. https://doi.org/10.1038/s41568-018-0016-5

Hu, J., Gauthier, J., Qian, P., Wilcox, E., & Levy, R. (2020). A systematic assessment of syntactic generalization in neural language models. In *Proceedings of ACL 2020* (pp. 1725-1744). https://doi.org/10.18653/v1/2020.acl-main.158

Johns, T. (1991). Should you be persuaded: Two samples of data-driven learning. *English Language Research Journal*, 4, 1-16.

Johns, T. (1998). Contexts: The background, development and trialing of a concordance-based CALL program. In A. Wichmann, S. Fligelstone, T. McEnery, & G. Knowles (Eds.), *Teaching and language corpora* (pp. 100-115). Longman.

Johnson, D. W., & Johnson, R. T. (1999). Making cooperative learning work. *Theory Into Practice*, 38(2), 67-73. https://doi.org/10.1080/00405849909543834

Joshi, P., Santy, S., Budhiraja, A., et al. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of ACL 2020* (pp. 6282-6293). https://doi.org/10.18653/v1/2020.acl-main.560

Jurafsky, D., & Martin, J. H. (2023). *Speech and language processing* (3rd ed. draft). Retrieved from https://web.stanford.edu/~jurafsky/slp3/

Kambanaros, M., & van Steenbrugge, W. (2006). Noun and verb processing in Greek progressive aphasia. *Brain and Language*, 97(2), 157-167. https://doi.org/10.1016/j.bandl.2005.10.001

Karasimos, A., & Makri, V. (2026, to be publ.). AI in English morphological processing: a GPT attempt or "misapproach"? *Selected papers of International Symposium on Theoretical and Applied Linguistics 26*. Aristotle University of Thessaloniki.

Karasimos, A., Makri, V., & Petropoulou, E. (2025). AI in German and Greek morphological nominal processing: an LLMs evaluation. Presented in *International Conference on Greek Linguistics*. 23-26 September 2025. Cambridge, UK: University of Cambridge.

Kasneci, E., Sessler, K., Küchemann, S., Bonnet, M., …, Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Keck, C. S., & Doarn, C. R. (2014). Telehealth technology applications in speech-language pathology. *Telemedicine and e-Health*, 20(7), 653-659. https://doi.org/10.1089/tmj.2013.0295

Khan, M. F. (2025). Syntactic Transformation in Large Language Models (LLMs): A Transformational Generative Grammar (TGG) Perspective. *Dibon Journal of Languages*, 1(1), 24–43. https://doi.org/10.64169/djl.23

Kihoza, P., Zlotnikova, I., Bada, J., & Kalegele, K. (2016). Classroom ICT integration in Tanzania: Opportunities and challenges from the perspectives of TPACK and SAMR models. *International Journal of Education and Development using ICT*, 12(1), 107-128.

Kintz, S., & Wright, H. H. (2018). Discourse measurement in aphasia research: Have we reached the tipping point? A review and analysis of discourse measures in aphasia research from 2011–2015. *Aphasiology*, 32, 13-35. https://doi.org/10.1080/02687038.2017.1398803

Koehler, M. J., & Mishra, P. (2009). What is technological pedagogical content knowledge? *Contemporary Issues in Technology and Teacher Education*, 9(1), 60-70.

Koleck, T. A., Dreisbach, C., Bourne, P. E., & Bakken, S. (2019). Natural language processing of symptoms documented in free-text narratives of electronic health records: A systematic review. *Journal of the American Medical Informatics Association*, 26(4), 364-379. https://doi.org/10.1093/jamia/ocy173

Kolk, H. (1995). A time-based approach to agrammatic production. *Brain and Language*, 50(3), 282-303. https://doi.org/10.1006/brln.1995.1049

Koutsikakis, J., Chalkidis, I., Malakasiotis, P., & Androutsopoulos, I. (2020). GREEK-BERT: The Greeks visiting Sesame Street. In *Proceedings of LREC 2020* (pp. 3149-3157). https://doi.org/10.1145/3411408.3411440

Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: A meta-analytic review. *Review of Educational Research*, 86(1), 42-78. https://doi.org/10.3102/00346543155814

Kulmizev, A., Bizzoni, Y., Nivre, J., et al. (2024). Do multilingual language models capture differing moral norms? *Computational Linguistics*, 50(1), 1-34. https://doi.org/10.48550/arXiv.2203.09904

Kuzman, T., Mozetič, I., & Ljubešić, N. (2023). *ChatGPT: Beginning of an end of manual linguistic data annotation? Use case of automatic genre identification*. arXiv preprint arXiv:2303.03953. https://doi.org/10.48550/arXiv.2303.03953

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge University Press.

Liddy, E. D. (2001). Natural language processing. In *Encyclopedia of library and information science* (2nd ed.). Marcel Dekker.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Havashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9), 1-35. https://doi.org/10.1145/3560815

Lombardi, M. M. (2007). Authentic learning for the 21st century: An overview. *Educause Learning Initiative*, 1(2007), 1-12.

Mita, M., Sakaguchi, K., Hagiwara, K., Mizumoto, T., Suzuki, J., & Kentaro, I. (2024). Towards Automated Document Revision: Grammatical Error Correction, Fluency Edits, and Beyond. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications* (BEA 2024) (pp. 251–265), Mexico City, Mexico. Association for Computational Linguistics.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed.). Lawrence Erlbaum Associates.

MacWhinney, B., & Wagner, J. (2010). Transcribing, searching and data sharing: The CLAN software and the TalkBank data repository. *Gesprachsforschung: Online-Zeitschrift zur verbalen Interaktion*, 11, 154-173.

MacWhinney, B., Fromm, D., Forbes, M., & Holland, A. (2011). AphasiaBank: Methods for studying discourse. *Aphasiology*, 25(11), 1286-1307. https://doi.org/10.1080/02687038.2011.589893

Manning, C. D. (2022). Human language understanding & reasoning. *Daedalus*, 151(2), 127-138. https://doi.org/10.1162/DAED_a_01905

Marini, A., Andreetta, S., Del Tin, S., & Carlomagno, S. (2011). A multi-level approach to the analysis of narrative language in aphasia. *Aphasiology*, 25(11), 1372-1392. https://doi.org/10.1080/02687038.2011.584690

McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.

Mercer, N., & Howe, C. (2012). Explaining the dialogic processes of teaching and learning: The value and potential of sociocultural theory. *Learning, Culture and Social Interaction*, 1(1), 12-21. https://doi.org/10.1016/j.lcsi.2012.03.001

Miller, J. F., & Iglesias, A. (2020). *Systematic analysis of language transcripts (SALT), research version 20*. Middleton, WI: Salt Software, LLC.

Min, S., Lyu, X., Holtzman, A., et al. (2022). Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of EMNLP 2022* (pp. 11048-11064). https://doi.org/10.18653/v1/2022.emnlp-main.759

Mishra, P., & Koehler, M. J. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. *Teachers College Record*, 108(6), 1017-1054. https://doi.org/10.1111/j.1467-9620.2006.0068

Nanousi, V., Masterson, J., Druks, J., & Atkinson, M. (2006). Interpretable vs. uninterpretable features: Evidence from six Greek-speaking agrammatic patients. *Journal of Neurolinguistics*, 19(3), 209-238. https://doi.org/10.1016/j.jneuroling.2005.11.003

National Research Council. (2000). *How people learn: Brain, mind, experience, and school: Expanded edition*. National Academies Press.

Nicholas, L. E., & Brookshire, R. H. (1993). A system for quantifying the informativeness and efficiency of the connected speech of adults with aphasia. *Journal of Speech, Language, and Hearing Research*, 36(2), 338-350. https://doi.org/10.1044/jshr.3602.338

Nivre, J., De Marneffe, M. C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., & Zeman, D. (2016). Universal dependencies v1: A multilingual treebank collection. In N. Calzolari, K. Choukri, H. Mazo, A. Moreno, T. Declerck, S. Goggi, M. Grobelnik, J. Odijk, S. Piperidis, B. Maegaard, & J. Mariani (Eds.), *Proceedings of the 10th International Conference on Language Resources and Evaluation*, LREC 2016 (pp. 1659-1666).

O'Donnell, M. (2008). The UAM CorpusTool: Software for corpus annotation and exploration. In *Proceedings of the XXVI Congreso de AESLA* (pp. 1433-1447).

Ouyang, L., Wu, J., Jiang, X., Almeida, D., …, & Lowe, R. (2022). Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, 35, 27730-27744. https://doi.org/10.48550/arXiv.2203.02155

Papathanasiou, I., Coppens, P., & Potagas, C. (Eds.). (2017). *Aphasia and related neurogenic communication disorders* (2nd ed.). Jones & Bartlett Learning.

Pea, R. D. (2004). The social and technological dimensions of scaffolding and related theoretical concepts for learning, education, and human activity. *Journal of the Learning Sciences*, 13(3), 423-451. https://doi.org/10.1207/s15327809jls1303_6

Pilehvar, M. T., & Camacho-Collados, J. (2019). *Embeddings in natural language processing: Theory and advances in vector representations of meaning*. Morgan & Claypool Publishers.

Prins, R., & Bastiaanse, R. (2004). Analysing the spontaneous speech of aphasic speakers. *Aphasiology*, 18(12), 1075-1091. https://doi.org/10.1080/02687030444000534

Prokopidis, P., & Papageorgiou, H. (2017). Universal Dependencies for Greek. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies* (pp. 102-106).

Puentedura, R. R. (2006). Transformation, technology, and education [Blog post]. Retrieved from http://hippasus.com/resources/tte/

Qin, C., Zhang, A., Zhang, Z., et al. (2023). Is ChatGPT a general-purpose natural language processing task solver? In *Proceedings of EMNLP 2023* (pp. 1339-1384). https://doi.org/10.18653/v1/2023.emnlp-main.85

Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine*, 380(14), 1347-1358. https://doi.org/10.1056/NEJMra1814259

Reeves, T. C., Herrington, J., & Oliver, R. (2002). Authentic activities and online learning. In *Quality conversations: Proceedings of the 25th HERDSA annual conference* (pp. 562-567).

Reppen, R. (2010). *Using corpora in the language classroom*. Cambridge University Press.

Roach, A., Schwartz, M. F., Martin, N., Grewal, R. S., & Brecher, A. (1996). The Philadelphia naming test: Scoring and rationale. *Clinical Aphasiology*, 24, 121-133.

Rochon, E., Saffran, E. M., Berndt, R. S., & Schwartz, M. F. (2000). Quantitative analysis of aphasic sentence production: Further development and new data. *Brain and Language*, 72(3), 193-218. https://doi.org/10.1006/brln.1999.2285

Römer, U. (2011). Corpus research applications in second language teaching. *Annual Review of Applied Linguistics*, 31, 205-225. https://doi.org/10.1017/S0267190511000055

Romrell, D., Kidder, L. C., & Wood, E. (2014). The SAMR model as a framework for evaluating mLearning. *Online Learning*, 18(2), 1-15. https://doi.org/10.24059/olj.v18i2.435

Royal College of Speech and Language Therapists. (2017). *Guidance on clinical education and supervision*. Royal College of Speech and Language Therapists.

Saffran, E. M., Berndt, R. S., & Schwartz, M. F. (1989). The quantitative analysis of agrammatic production: Procedure and data. *Brain and Language*, 37(3), 440-479. https://doi.org/10.1016/0093-934x(89)90030-8

Salomon, G., Perkins, D. N., & Globerson, T. (1991). Partners in cognition: Extending human intelligence with intelligent technologies. *Educational Researcher*, 20(3), 2-9. https://doi.org/10.3102/0013189X02000300

Schmid, M., Brianza, E., & Petko, D. (2020). Developing a short assessment instrument for Technological Pedagogical Content Knowledge (TPACK.xs) and comparing the factor structure of an integrative and a transformative model. *Computers & Education*, 157, 103967. https://doi.org/10.1016/j.compedu.2020.103967

Schwartz, M. F., Kimberg, D. Y., Walker, G. M., et al. (2009). Anterior temporal involvement in semantic word retrieval: VLSM evidence from aphasia. *Brain*, 132(12), 3411–3427. https://doi.org/10.1093/brain/awp284

Sclar, M., Choi, Y., Tsvetkov, Y., & Suhr, A. (2023). *Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting*. arXiv preprint arXiv:2310.11324. https://doi.org/10.48550/arXiv.2310.11324

Seabold, S., & Perktold, J. (2010). statsmodels: Econometric and statistical modeling with Python. In *Proceedings of the 9th Python in Science Conference* (pp. 56-61).

Shen, D., Wu, G., & Suk, H. I. (2017). Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, 19, 221-248. https://doi.org/10.1146/annurev-bioeng-071516-044442

Shen, C., Cheng, L., Nguyen, X-N., You, Y., & Bing, L. (2023). Large Language Models are Not Yet Human-Level Evaluators for Abstractive Summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (4215–4233). https://doi.org/10.18653/v1/2023.findings-emnlp.278

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, 15(2), 4-14. https://doi.org/10.2307/1175860

Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review*, 57(1), 1-23.

Simmons-Mackie, N., Raymer, A., Armstrong, E., Holland, A., & Cherney, L. R. (2010). Communication partner training in aphasia: A systematic review. *Archives of Physical Medicine and Rehabilitation*, *91*(12), 1814–1837. https://doi.org/10.1016/j.apmr.2010.08.026

Sloetjes, H., & Wittenburg, P. (2008). Annotation by category: ELAN and ISO DCR. In *Proceedings of the 6th International Conference on Language Resources and Evaluation* (pp. 816-820).

Stark, B. C., Meinert, K., Urena, K., Oeding, G., Fromm, D., & MacWhinney, B. (2025). Introducing the NEURAL Research Lab Data Set for Studies of Discourse and Gesture in Aphasia and Cognitively Healthy Aging Adults. *Journal of speech, language, and hearing research* (JSLHR), 68(11), 5543–5556. https://doi.org/10.1044/2025_JSLHR-24-00732

Stephany, U. (1997). The acquisition of Greek. In D. I. Slobin (Ed.), *The crosslinguistic study of language acquisition* Vol. 4 (pp. 183-333). Lawrence Erlbaum.

Sullivan, M., Kelly, A., & McLaughlan, P. (2023). ChatGPT in higher education: Considerations for academic integrity and student learning. *Journal of Applied Learning & Teaching*, 6(1), 1-10. https://doi.org/10.37074/jalt.2023.6.1.17

Thompson, C. K., Cho, S., Hsu, C. J., Wieneke, C., Rademaker, A., Weitner, B.B., Mesulam, M-M. & Weintraub, S. (2012). Dissociations between fluency and agrammatism in primary progressive aphasia. *Aphasiology*, 26(1), 20-43. https://doi.org/10.1080/02687038.2011.584691

Tucker, J. K. (2012). Perspectives of speech-language pathologists on the use of telepractice in schools: The qualitative view. *International Journal of Telerehabilitation*, 4(2), 47-60. https://doi.org/10.5195/ijt.2012.6102

Üstün, A., Bisazza, A., Bouma, G., KO, W-Y.,... & Hooker, S. (2024). *Aya model: An instruction finetuned open-access multilingual language model*. arXiv preprint arXiv:2402.07827. https://doi.org/10.48550/arXiv.2402.07827

van de Pol, J., Volman, M., & Beishuizen, J. (2010). Scaffolding in teacher-student interaction: A decade of research. *Educational Psychology Review*, 22(3), 271-296.

Varlokosta S., Stamouli S., Karasimos A., Markopoulos G., Kakavoulia M., Nerantzini M., Pantoula A., V. Fyndanis, Economou, A. & A. Protopapas (2016). A Greek Corpus of Aphasic Discourse: Collection, Transcription and Annotation Specifications. In *LREC 2016 Proceedings Workshop RAPID2016*, pp. 14-21. ISSN 2522-2686.

Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, 30, 5998-6008.

Voogt, J., Fisser, P., Pareja Roblin, N., Tondeur, J., & van Braak, J. (2013). Technological pedagogical content knowledge—A review of the literature. *Journal of Computer Assisted Learning*, 29(2), 109-121. https://doi.org/10.1111/j.1365-2729.2012.00487.x

Vyatkina, N. (2020). Corpora as open educational resources for language teaching. *Foreign Language Annals*, 53(2), 359-370. https://doi.org/10.1111/flan.12464

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Wang, Y., Wang, L., Rastegar-Mojarad, M., Moon, S., Shen, F., Afzal, N., Liu, S., Zeng, Y., Mehrabi, S., Sohn, S., & Liu, H. (2018). Clinical information extraction applications: A literature review. Journal of biomedical informatics, 77, 34–49. https://doi.org/10.1016/j.jbi.2017.11.011

Warstadt, A., & Bowman, S. R. (2022). What artificial neural networks can tell us about human language acquisition. In J. Sprouse (Ed.), *Oxford handbook of experimental syntax* (pp. 1-44). Oxford University Press. https://doi.org/10.1201/9781003205388-2

Webster, J., Franklin, S., & Howard, D. (2007). An analysis of thematic and phrasal structure in people with aphasia: What more can we learn from the story of Cinderella? *Journal of Neurolinguistics*, 20(5), 363-394. https://doi.org/10.1016/j.jneuroling.2007.02.002

Weintrop, D., Beheshti, E., Horn, M., Orton, K., Jona, K., Trouille, L., & Wilensky, U. (2016). Defining computational thinking for mathematics and science classrooms. Journal of Science Education and Technology, 25(1), 127–147. https://doi.org/10.1007/s10956-015-9581-5

Wenger, E. (1998). *Communities of practice: Learning, meaning, and identity*. Cambridge University Press.

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Einashar, A., Spenser-Smith, J., & Schmidt, D. (2023). *A prompt pattern catalog to enhance prompt engineering with ChatGPT*. arXiv preprint arXiv:2302.11382. https://doi.org/10.48550/arXiv.2302.11382

M. Wilson, S., Eriksson, D. K., Brandt, T. H., Schneck, S. M., Lucanie, J. M., Burchfield, A. S., Charney, S., Quillen, I. A., de Riesthal, M., Kirshner, H. S., Beeson, P. M., Ritter, L., & Kidwell, C. S. (2019). Patterns of recovery from aphasia (Version 1). *ASHA journals*. https://doi.org/10.23641/asha.7811876.v1

Wing, J. M. (2006). Computational thinking. *Communications of the ACM*, 49(3), 33-35. https://doi.org/10.1145/1118178.111821

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. https://doi.org/10.1111/j.1469-7610.1976.tb00381.x

Woolf, B. P., Lane, H. C., Chaudhri, V. K., & Kolodner, J. L. (2013). AI grand challenges for education. *AI Magazine*, 34(4), 66-84. https://doi.org/10.1609/aimag.v34i4.2490

World Health Organization. (2001). *International classification of functioning, disability and health (ICF)*. World Health Organization.

Wu, Z., Qiu, L., Ross, A., Akyürek, E., Chen, B., Wang, B., Kim, N., Andreas, J. & Kim, Y. (2024). Reasoning or Reciting? Exploring the Capabilities and Limitations of Language Models Through Counterfactual Tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1819–1862), Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.102

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education. *International Journal of Educational Technology in Higher Education*, 16(1), 1-27. https://doi.org/10.1186/s41239-019-0171-0

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., … J.-R. Wen (2025). *A survey of large language models. arXiv preprint.* arXiv:2303.18223. https://doi.org/10.48550/arXiv.2303.18223

Appendix A: Educational Materials

Automated and Manual Linguistic Annotation: LLM-Powered Corpus Analysis with CACLA	
Athanasios Karasimos	
Level of English Proficiency	This activity is for students at C1-C2 CEFR level of English proficiency (advanced undergraduate or graduate students in linguistics, computational linguistics, or NLP programs)
Course Area	This activity is designed for courses in *Introduction to Computational Linguistics*,*Natural Language Processing and Computational Techniques*, and *Corpus Linguistics and Language Teaching*.
Target Skill(s)	Corpus linguistics methodology, computational annotation techniques, morphosyntactic analysis, prompt engineering for NLP tasks, critical evaluation of LLM capabilities, collaborative computational research
Time Frame	The estimated timeframe for this activity is approximately two 150-minute lectures (5 hours total instruction time plus homework assignments)
Apps/Software needed	- LLM Software (Claude, GPT, Gemini, DeepSeek, Grok) - Web browser (Chrome, Firefox, Safari) - ELAN software (for comparison with traditional annotation tools) - Google Colab or Jupyter Notebook (optional, for advanced students) - Text editor with UTF-8 support (e.g., NotePad++, Sublime Text) - Spreadsheet software (Microsoft Excel, Google Sheets) for data compilation and statistical analysis - (Optional) Python environment with NLTK, spaCy, or similar NLP libraries for comparative analysis
Materials	- Selected transcripts from CACLA Corpora in plain text format - Linguistic annotation scheme reference guide (morphological, syntactic, lexical levels - parts of annotation schema template) - LLM prompt templates for computational annotation tasks - Gold standard annotated corpus samples for validation - Comparative analysis worksheet for LLM vs. rule-based/statistical methods - Assessment rubric for computational linguistics report - Research articles on corpus annotation methods and LLM evaluation - Greek linguistic reference materials (morphology, syntax overview, aphasia)
Overview	This lesson plan introduces Computational linguistics students to modern LLM-based approaches for linguistic corpus annotation and analysis. Using the CACLA corpora of aphasic speech as a case study, students will explore how Large Language Models can be leveraged for automated morphological, lexical, and syntactic annotation tasks. Through systematic prompt engineering, validation against gold standards, and comparative analysis with traditional computational methods, students develop critical skills in evaluating AI-driven NLP tools while understanding both their

	transformative potential and inherent limitations. The activity emphasizes the computational linguistics perspective: accuracy metrics, annotation consistency, linguistic feature extraction, and the methodological rigor required when deploying LLMs for serious linguistic research.
Aims of the Activities	At the end of this lesson plan/lecture, students will be able to: - Design and implement LLM-based pipelines for multi-level linguistic annotation- Apply prompt engineering principles to optimize linguistic analysis tasks - Extract quantitative linguistic measures from annotated corpora (TTR, MLU, morphological complexity, syntactic diversity) - Validate automated annotations systematically using gold standard comparisons - Calculate and interpret standard evaluation metrics (precision, recall, F1-score, inter-annotator agreement) - Compare LLM performance with traditional rule-based and statistical NLP methods - Identify linguistic phenomena that challenge current LLM architectures- Document computational workflows with reproducible methodologies - Critically evaluate the suitability of LLMs for different annotation tasks - Understand the special challenges of analyzing atypical (aphasic) language data - Collaborate effectively on computational linguistics research projects
Prerequisites	- Basic knowledge of linguistic analysis (morphology, syntax, semantics) - Completion of introductory corpus linguistics course or equivalent - Familiarity with Greek language structure - Basic understanding of NLP concepts (tokenization, POS tagging, parsing) - Familiarity with statistical concepts (mean, standard deviation, correlation) - Creating a free LLM account - Ability to read and understand linguistic research papers in English - (Recommended) Prior exposure to annotation tools like ELAN, UAM CorpusTool, or similar - (Optional) Experience with at least one programming language (Python preferred) at basic level
Lecture Structure (all the stages)	
Preparation	The lecturer prepared: - Curated selection of CACLA transcripts (8-12 samples, 200-400 words each, representing varied linguistic complexity - mainly from Greek TV Series scripts) - Gold standard annotations at three levels: morphological, lexical, syntactic (extracted human annotations from ELAN) - Comprehensive prompt template library organized by linguistic annotation task - PowerPoint presentation on: - CACLA corpus structure and annotation scheme - LLM capabilities and limitations for linguistic analysis - Evaluation metrics for NLP systems - Video demonstration/tutorial (10-12 minutes) showing complete LLM

	annotation workflow - Detailed annotation guidelines document with linguistic examples - Comparative analysis worksheet templates - Assessment rubric for computational linguistics report - Research paper reading list on corpus annotation and LLM evaluation (Zotero, Mendeley) - Shared computational workspace (Google Drive, Microsoft One Drive) for code and data sharing - (Optional) Baseline annotation outputs from traditional NLP tools (spaCy) for comparison - Statistical analysis template spreadsheet for calculating evaluation metrics
Process	**Step 1 (Lecture 1): Introduction to Corpus Linguistics Foundations and CACLA Project (90 minutes)** - Overview presentation of corpus linguistics methodology: corpus design principles, annotation schemes, representativeness, and research applications - Introduce the CACLA corpora as a computational linguistics case study: - Corpus composition and structure - Transcription conventions and metadata - Linguistic characteristics of atypical language data - Annotation challenges specific to clinical corpora - Present the linguistic annotation task framework: - Morphological annotation: lemmatization, POS tagging, morphological features (case, number, gender, tense, aspect, mood for Greek) - Lexical annotation: word frequency, lexical diversity measures, lexical categories - Syntactic annotation: phrase structure, dependency relations, grammaticality judgments - Discuss traditional computational approaches to these tasks (recap): - Rule-based systems (advantages and limitations) - Statistical models (CRF, HMM for sequence tagging) - Neural approaches (BiLSTM, transformer-based models) - Divide students into groups of 3-4. Each group receives: - Two CACLA transcripts (one simpler, one more complex linguistically) - Digital and printed versions - Annotation guidelines document - Groups conduct manual linguistic analysis of the simpler transcript (30 minutes- *ideally they use ELAN and an annotation template*): - Identify and annotate 15-20 words with full morphological analysis - Note interesting linguistic phenomena (errors, unusual structures, ambiguities) that are included in the hierarchal annotation schema - Discuss annotation challenges and ambiguous cases - Brief group presentations (3-4 minutes each): What linguistic patterns did you observe? What annotation challenges and issues arose? How consistent were your annotations within the group? - Class discussion: Introduce the concept of inter-annotator agreement and why consistency matters in computational linguistics (see Varlokosta et al.,

2016).

Step 2 (Lecture 1): LLM-Based Annotation Pipeline Development (75 minutes)

- Present comprehensive overview of Large Language Models for linguistic analysis:
    - LLM architecture basics (transformer models, attention mechanisms at appropriate level)
    - How LLMs process and generate language
    - Documented linguistic capabilities and limitations
    - Comparison with traditional NLP models
- Live demonstration of LLM annotation workflow using Claude/GPT:
    - Uploading corpus transcript
    - Designing effective prompts for morphological annotation
    - Requesting structured output formats (JSON, TSV, tables)
    - Iterative prompt refinement based on output quality
    - Extracting and processing results
- Present principles of prompt engineering for linguistic tasks:
    - Clarity and specificity in linguistic terminology
    - Providing examples (from zero-shot to few-shot prompting)
    - Specifying output format and structure
    - Handling ambiguity and requesting confidence indicators
    - Chaining prompts for complex multi-step analyses
    - Prompt templates vs. dynamic prompt generation
- Introduce the prompt template library with examples for:
    - Morphological annotation (lemmatization, POS tagging, feature extraction)
    - Lexical analysis (frequency, diversity, semantic fields)
    - Syntactic annotation (constituency parsing, dependency relations)
- Quantitative measure extraction (MLU, TTR, complexity metrics)
- Error detection and linguistic anomaly identification
- Groups begin computational annotation project:
    - Each group works with both assigned transcripts
    - Divide annotation responsibilities among group members (each takes one linguistic level)
    - Access LLM and upload transcripts
    - Use and adapt provided prompt templates
    - Document all prompts used and modifications made
    - Generate initial automated annotations
    - Compile outputs in structured format (spreadsheet or JSON)
- Groups share preliminary findings: *What worked? What failed? How did you refine prompts?*
- (optional) Troubleshooting session: Addressing common issues (encoding problems, inconsistent output formats, Greek language handling)

Step 3 (Lecture 2): Validation, Evaluation Metrics, and Comparative Analysis (90 minutes)

- Present rigorous evaluation methodology for NLP systems:
    - Gold standard creation and its importance

	- Evaluation metrics: precision, recall, F1-score, accuracy - Confusion matrices for error analysis - Statistical significance testing - Error categorization and analysis - Distribute gold standard annotations for the transcripts students have been analyzing - Systematic validation exercise (60 minutes): - Quantitative evaluation: *- Groups compare their LLM-generated annotations with gold standards* *- Calculate precision, recall, and F1-score for some annotation types* *- Create confusion matrices for POS tagging* *- Identify systematic error patterns* *- Use provided spreadsheet templates for metric calculation* - Qualitative error analysis: *- Categorize errors by type (linguistic level, error source)* *- Identify which linguistic phenomena the LLM handles well/poorly* *- Analyze whether errors correlate with specific Greek morphological features* *- Note whether atypical language characteristics affect accuracy* *- Examine false positives vs. false negatives* - Comparative analysis: *- Compare LLM performance across different annotation tasks* *- Compare simple vs. complex transcripts* *- Analyze time efficiency: manual vs. automated annotation* *- Consider accuracy-speed trade-offs* *- Introduce advanced measurement extraction:* *- Design prompts for calculating linguistic complexity metrics:* *- Mean Length of Utterance (MLU)* *- Type-Token Ratio (TTR) and variants (MTLD, MATTR)* *- Morphological complexity indices* *- Syntactic complexity measures (subordination index, dependency distance)* - Groups extract these measures from their annotated corpora - Compile results in shared spreadsheet for cross-group comparison - Groups complete comprehensive comparative analysis worksheets: - Strengths of LLM approach for each annotation type - Specific limitations discovered with examples - Computational efficiency analysis - Recommendations for optimal use cases - Suggestions for hybrid approaches (LLM + rule-based/statistical methods) **Step 4 (Lecture 2): Advanced Applications, Reproducibility, and Research Presentations (75 minutes)**

	- Brief teacher presentation on advanced computational linguistics applications: - Batch processing multiple corpus files efficiently - Using Claude's API for programmatic access (Python examples) - Building custom annotation artifacts with visualization - Developing hybrid pipelines (LLM + traditional NLP tools) - Reproducibility best practices in computational linguistics research - Version control for corpus annotation projects - Optional advanced exercise for interested students: - Use LLM to generate Python code for batch annotation processing - Create data visualization scripts for linguistic feature distributions - Groups finalize comprehensive computational linguistics research reports focusing on: - Methodology section: Clear description of corpus, annotation scheme, LLM configuration, prompts used, evaluation procedures - Results section: Quantitative evaluation metrics with tables and figures, error analysis with examples, linguistic measure extraction results - Comparative analysis: LLM vs. traditional methods, accuracy-efficiency trade-offs, task-specific performance differences - Discussion: Linguistic phenomena that challenge LLMs, implications for computational linguistics research, recommendations for future work - Reproducibility: All prompts documented, workflow clearly described, data and code availability stated - Each group member presents their specialized section - Class discussion synthesizing findings across all groups: - Which annotation tasks are most suitable for LLM automation? - What linguistic features consistently challenge LLMs? - How does atypical language affect LLM performance? - What role should LLMs play in computational linguistics research? - Reflection on computational linguistics methodology: - Balancing automation with linguistic rigor - Importance of validation and transparency - Future directions for LLM applications in corpus linguistics - Ethical considerations in computational analysis of clinical data
Variations	For different computational backgrounds: - *Advanced Computer Science students*: Emphasize algorithmic aspects, include API usage, develop Python scripts for batch processing, implement statistical significance testing, compare with transformer models like BERT/GPT - *Linguistics students*: Focus on linguistic analysis quality, reduce programming components, emphasize qualitative error analysis, connect to theoretical linguistics For different linguistic focus: - *Morphology-focus*: Deep dive into Greek morphological complexity, morpheme segmentation, allomorphy, paradigm analysis - *Syntax-focus*: Focus on parsing accuracy, dependency relations, constituency structures, grammaticality judgments

	- *Semantics-focus*: Lexical semantics annotation, semantic role labeling, meaning representations Alternative corpora and applications: - Replace CACLA with other corpora: child language acquisition, language disabilities data, learner corpora - Cross-linguistic comparison: Apply same methodology to parallel corpora in multiple languages Technology variations: - Compare multiple LLMs (Claude vs. GPT-5 vs. Gemini) for same tasks - Use specialized linguistic annotation platforms (WebAnno, Sketch Engline, CatMa) alongside LLMs
Troubleshooting Tips	**Technical Issues:** - *LLM access problems*: Have backup accounts ready by providing alternative LLM options - *Greek text encoding*: Ensure UTF-8 encoding throughout; test copy-paste vs. file upload; provide pre-processed clean text files - *Response length limitations*: Break long transcripts into smaller chunks; use multi-turn conversations; teach students to request continuation **LLM-Specific Challenges:** - *Hallucinated annotations*: Train students to always validate against source text and do not easily accept the output. - *Inconsistent annotation across examples*: Design prompts emphasizing consistency. - *Greek morphological errors*: Provide LLM with explicit Greek morphology reference in prompt; use few-shot examples. - *Ambiguity handling*: Request confidence scores or multiple analyses; teach students to identify genuinely ambiguous cases vs. LLM errors **Pedagogical Challenges:** - *Over-reliance on automation*: Require manual annotation baseline for comparison; grade based on critical evaluation, not just LLM output; emphasize validation importance - *Insufficient linguistic knowledge*: Provide supplementary materials on Greek morphosyntax; pair stronger/weaker students strategically. - *Time management issues*: Provide realistic time estimates for each task; prioritize core activities if running behind. - *Statistical concepts confusion*: Provide worked examples of metric calculations; offer spreadsheet templates with formulas; explain concepts with linguistic examples rather than pure math **Computational Challenges:** - *Data format issues*: Provide clear format specifications and examples; offer conversion scripts; validate formats before processing - *Evaluation metric calculation errors*: Provide spreadsheet templates with built-in formulas; show step-by-step calculations; offer troubleshooting session

Links	Primary Tools: LLMs: GPT https://chatgpt.com/ | Gemini https://gemini.google.com/app Claude https://claude.ai/ | DeepSeek https://www.deepseek.com/en | Grok https://grok.com/ CACLA Corpus Resources: https://enl.auth.gr/cacla/ Corpus Linguistics Resources: Introduction to Corpus Linguistics: https://www.corpuslinguis.org Lancaster University Corpus Linguistics Page: https://www.lancaster.ac.uk/linguistics/about/research/corpus-linguistics/ Corpus Linguistics Tutorial (YouTube): [relevant playlist] Text Encoding Initiative (TEI) Guidelines: https://tei-c.org/guidelines/ NLP and Computational Tools: ELAN Annotation Software: https://archive.mpi.nl/tla/elan Natural Language Toolkit (NLTK): https://www.nltk.org spaCy NLP Library: https://spacy.io UAM CorpusTool: http://www.corpustool.com AntConc: https://www.laurenceanthony.net/software/antconc/ Evaluation and Metrics: SciKit-Learn Metrics Documentation: https://scikit-learn.org/stable/modules/model_evaluation.html LLM and Prompt Engineering: Prompt Engineering Guide: https://www.promptingguide.ai OpenAI Prompt Engineering: https://platform.openai.com/docs/guides/prompt-engineering Anthropic Prompt Engineering Guide: https://docs.anthropic.com/en/docs/build-with-claude/prompt-engineering Few-Shot Learning with LLMs: https://www.coursera.org/learn/zero-shot--few-shot-learning-master-ai-with-minimal-data
Other Information	**Copyright and Ethics:** - All CACLA corpus materials used with appropriate authors/creators permissions and ethical approvals - LLM-assisted analysis must be clearly labeled in all reports and presentations - Students must acknowledge AI tools used: "Linguistic annotations were generated using LLM name (Company name) and validated against gold standard annotations" **Accessibility Considerations:** - Provide transcripts in multiple formats (plain text, annotated, screen-reader friendly) - Ensure all video materials have captions - Allow flexible group composition to accommodate diverse needs - Provide extended time for students who require it

	Future Development: - Activity template will be iteratively refined based on instructor feedback - Additional expansion of the CACLA corpora and linguistic phenomena may be incorporated - Evaluation rubrics may be adjusted for different course contexts

Athanasios Karasimos (akarasimos@enl.auth.gr) is an Assistant Professor in Computational Linguistics, Aristotle University of Thessaloniki, School of English. He is a graduate of the Department of Philology, University of Patras. He holds two European Masters in Speech and Language Processing (one of them at the University of Edinburgh) and his doctoral dissertation is in Computational Morphology. He participated in several research projects on Modern Greek dialects, corpora, aphasic speech, Digital Humanities, and training of English language teachers. He was a postdoctoral research fellow funded by IKY. He worked as an Adjunct Lecturer at HOU, AUTh, and NKUA teaching Educational Technology, Research Methodology, Computational Linguistics and Corpus Linguistics. He is a researcher in the national infrastructure for Digital Humanities DARIAH-GR / DYAS (Academy of Athens). His research interests focus on Computational Linguistics and machine learning, the use of corpora, education technology, and integrating video and board games into language teaching and learning.

Evangelia–Antonia Efstratiadou (e.efstratiadou@go.uop.gr) is an Assistant Professor in Speech and Language Therapy at the University of the Peloponnese. She is a graduate of the Technological Educational Institute of Western Greece with a degree in Speech and Language Therapy. She holds an MSc in Human Communication and a PhD in Language and Communication Sciences from City, University of London, where her doctoral research focused on aphasia therapy. She has over fifteen years of clinical experience in paediatric and adult neurogenic communication disorders and has worked in rehabilitation and hospital settings providing assessment, intervention, and clinical supervision. She has participated in large-scale national and European research projects, including the THALES Project, the CACLA Project, and MapCosmos. She co-leads national initiatives on aphasia assessment and implementation science, contributing to the development of LexiGrAph, a Greek lexical and grammatical aphasia assessment tool, and to the adaptation of KomTil for Greek clinical contexts. Her work has been presented at major international conferences such as IARC and ICPLA. Her research interests focus on aphasia rehabilitation, evidence-based intervention methods (such as Semantic Feature Analysis and Mapping Therapy), psycholinguistic mechanisms underlying language processing, lexical and grammatical assessment in Greek, communication partner training, implementation science in speech and language therapy, qualitative and quantitative analysis of aphasic speech, and the translation of research evidence into clinical practice. She also has extensive experience in training and mentoring healthcare professionals working with individuals with acquired communication disorders.

Dr. Christos Papatzalas (cpapatz@upatras.gr) is an Assistant Professor in Speech-Language Pathology: Communication Disorders, Department of Speech and Language Therapy, University of Patras. He holds a PhD in Clinical Speech-Language Pathology from the University of Thessaly (Faculty of Medicine, Neurosurgery Clinic), focusing on language assessment during awake brain surgery,

completed with a three-year scholarship from the Greek State Scholarships Foundation (IKY). He also holds a BSc in Speech and Language Therapy and an MSc in Linguistics from the University of Patras. His research focuses on communication disorders (and bilingualism) with emphasis on aphasia, tumor-related language disorders, and intraoperative language mapping. He has participated in multiple research projects and has published in international peer-reviewed journals and conference proceedings.

Dr. Ilias Papathanassiou (ipapatha@upatras.gr) is a Professor of Speech and language Therapy, at the Department of Speech and Language Therapy, University of Patras. He has almost 30 years of experience in the area of rehabilitation voice, neurogenic communication and swallowing disorders. He was awarded his PhD from the Institute of Neurology, University College London, University of London and has done postdoctoral studies at the Department of Speech and Language Therapy at University of Queensland, Australia. He has published books (Acquired Neurogenic Communication Diosrders, 2000; The Sciences of Aphasia, 2002; Aphasia and Related Neurogenic Communication Disorders, 2012, 2016, 2022), has published dozens of scientific articles and studies and has over a hundred international participations with presentations in an international peer reviewed conferencesHe has served as a member of the editorial board of the scientific journals Aphasiology, Communication Disorders Quarterly, Clinical Archives of Communication Disorders, ASHA perspectives SIG17, and has been responsible for the book reviews of the International Journal of Language and Communication Disorders and Folia Phoniatrica et logopedica. Dr. Papathanasiou has received awards in both Great Britain and America. In 2008, the Royal College of Speech Therapists in Great Britain awarded him the honorary title of. In 2014, the American Speech Language Hearing Association (ASHA) was awarded the honorary title of Fellow (the Highest Award for Speech and Language Therapist in the USA) for its contribution to research, teaching and international organizations of the profession. He is the 2020 recipient of the ASHA Certificate of Recognition for Outstanding Contributions in International Achievement.

Research Papers in Language Teaching and Learning
Vol. 16, No. 1, March 2026, 141-155
ISSN: 1792-1244
Available online at http://rpltl.eap.gr

Greek Dialogues in the Banking Domain: Large Language Models, Data Evaluation, and Pedagogical Applications

Alexandra Fiotaki

Artificial intelligence enables the generation of domain-specific dialogues that can serve as teaching resources in language education. This study focuses on the Greek banking domain, a communicative context requiring formal registers, polite requests, and specialized financial vocabulary. Given Greek's rich morphology, such dialogues demand heightened grammatical and pragmatic accuracy. Three large language models (Krikri, Gemini, and ChatGPT) were prompted to produce dialogues for typical banking scenarios such as opening an account or resolving service issues. The comparative analysis evaluates morphosyntactic accuracy, register appropriateness, pragmatic naturalness, dialogue resolution, and domain-specific terminology. Results reveal distinct linguistic profiles: ChatGPT demonstrates the highest dialogue resolution rate with compact output, Gemini produces the most natural and empathetically rich interactions despite greater verbosity, while Krikri exhibits higher lexical diversity but is constrained by shorter dialogue length and a significant rate of unresolved interactions. Patterns of English influence on Greek output, including literal translations and untranslated loanwords, were also identified. Pedagogically, the study proposes an annotation-led framework where AI-generated dialogues serve as classroom materials for role playing, error spotting, and targeted linguistic reflection, highlighting their potential to enrich language learning through meaningful, context-sensitive communication.

Keywords: Large Language Models, data evaluation, pedagogical application, dialogue systems, banking domain

---

1. Introduction

Teaching in professional and academic contexts requires more than transmitting abstract knowledge; it also involves preparing learners to navigate real-world communicative situations. Teachers often struggle to find teaching resources that are realistic and domain-specific enough for specialized interactions, such as those occurring in financial, medical, or legal settings. Learners in these contexts need to practice not only subject knowledge but also the communicative strategies and vocabulary that characterize professional exchanges.

Recent developments in Artificial Intelligence (AI) have further expanded the pedagogical possibilities available to educators (Williamson & Eynon, 2020; Godwin-Jones, 2021; Kohnke et al., 2023). While traditional textbooks offer structured grammatical progression, they often fail to capture the dynamic and context-specific nature of real-world communicative encounters. AI-enhanced approaches can support student engagement and self-regulation while also improving teaching effectiveness and promoting more interactive forms of communication (Seo et al., 2021; Ng et al., 2023).

Within this landscape, large language models (LLMs) offer promising applications for language teaching (Baskara & Mukarto, 2023; Jeon & Lee, 2023), as they can generate domain-specific dialogues that mirror authentic professional interactions, providing educators with flexible and contextually rich materials. However, the effectiveness of AI-generated dialogues requires critical examination: while LLMs can produce realistic scenarios at scale, their accuracy, register appropriateness, and pragmatic naturalness vary significantly across models. This means some outputs may suit direct classroom use, while others may require curation or serve as pedagogical tools for critical analysis.

Banking offers an ideal setting for pedagogical work. Research in Languages for Specific Purposes has long acknowledged that banking interactions require specialized communicative competence, including formal registers, specialized terminology, and culturally appropriate politeness strategies. Common banking scenarios require learners to request information, compare options, and negotiate outcomes while balancing technical precision with interpersonal sensitivity. For Greek, these challenges are heightened by the language's rich morphology, its formal-informal register contrasts, and the scarcity of specialized Greek teaching materials. Despite their pedagogical value, authentic banking dialogues are rarely available in teaching resources due to their sensitive content.

This paper explores the potential and limitations of using LLM-generated dialogues for teaching Greek in the banking domain. Its primary contribution is a reusable dataset of Greek banking dialogues, accompanied by an evaluation of their linguistic quality. Notably, the corpus is entirely synthetic, generated through prompted interactions with language models, making the findings prompt-dependent and model-dependent, not directly generalizable to authentic banking discourse. We first outline the dialogue generation methodology and analysis framework, then present a detailed analysis focusing on morphosyntactic accuracy, register appropriateness, pragmatic effectiveness, and handling of specialized vocabulary. Building on these findings, we propose pedagogical applications for both curated and non-curated dialogues and reflect on the broader implications of integrating AI-generated materials into language teaching practice.

2. Methodology: Dialogue Generation and Analysis Framework

The methodology for this study was designed to generate and evaluate LLM-generated Greek banking dialogues for pedagogical use. The research was structured in three phases: (1) Corpus Design and Prompt Engineering (Section 2.1), where banking scenarios and prompting strategies were defined; (2) Dialogue Generation (Section 2.2), where the methodology of dialogue generation is presented; and (3) Evaluation Framework (Section 2.3), where the methodology used to evaluate the generated dialogues is presented. Together, these phases ensure that the dataset is not only diverse and representative but also systematically assessed for its linguistic quality and pedagogical suitability.

2.1 Corpus Design and Prompt Engineering

The corpus was designed to encompass dialogues reflecting a wide range of authentic banking scenarios, ensuring representation of the linguistic, pragmatic, and stylistic diversity typical of real customer service interactions. To ground the design in authentic communicative situations, the

prompts were informed by real example data provided by *Omilia* - Conversational Intelligence[1]. All sensitive user information was removed from the reference data prior to use, ensuring compliance with data privacy standards while preserving the structural and linguistic features of genuine banking interactions. The resulting corpus includes communicative situations such as financial transactions, problem-solving exchanges, service inquiries, and investment - related discussions.

To achieve this level of variation and realism, a detailed prompt was developed to guide the generation of dialogues by LLMs. The prompt's specifications were designed with particular attention to the structural, linguistic, and interactional dimensions that underpin the dataset, aiming to simulate authentic and natural exchanges between users and agents in Greek telephone banking contexts. The prompt mandates a standardized "User" and "Agent" format, excluding all metadata, and specifies dialogue lengths between 8 and 30 turns to encompass transactional, problem-solving, inquiry, and investment scenarios. It further defines parameters such as participant roles, degree of formality, emotional language, and language proficiency.

The design of the prompt draws on established principles of prompt engineering, which emphasize that the structure and specificity of a prompt directly influence the quality, relevance, and accuracy of generated materials (Kohnke et al., 2023). Prompting is not a simple act of inputting a request; rather, it is a deliberate communicative strategy that involves framing instructions, constraints, and contextual cues to elicit targeted and pedagogically valuable output (White et al., 2023). Techniques such as zero-shot, few-shot, and chain-of-thought prompting (Brown et al., 2020; Wei et al., 2022) offer educators a range of strategies for generating linguistically precise and domain-specific content.

The prompt used for corpus generation combined multiple engineering techniques to optimize control over the LLM's output and ensure pedagogical alignment with language learning goals in banking (Geroimenko, 2025). The design incorporated persona priming to assign a domain-specific expert role, strict output formatting rules, and negative constraints to prevent unwanted metadata (Brown et al., 2020), along with rule-based content quotas for diversity and avoiding repetitive scenarios (Wei et al., 2022). It also included explicit instructions for fine-grained speech features such as fillers, pauses, and self-corrections to simulate natural human speech disfluencies and enhance authenticity.

Collectively, these features reflect a balanced integration of prompt engineering techniques, domain coverage, stylistic conditioning, and structural control to produce a linguistically varied and pedagogically purposeful corpus of banking dialogues[2].

2.2 Dialogue Creation

The second phase of the methodology involved the systematic generation of the dialogue corpus. The prompt designed was deployed across the LLMs: OpenAI's ChatGPT-4.1, Google's Gemini 2.5, and ILSP's Krikri-8B-Instruct. ChatGPT and Gemini were selected as leading general-purpose multilingual models with strong generation capabilities, while Krikri, a language model developed specifically for Greek, was included to provide a language-specific baseline and to explore whether a model with dedicated Greek training data would produce more linguistically accurate or natural outputs.

---

[1] This study was conducted with the kind support of *Omilia*, which provided access to the necessary resources and infrastructure.

[2] To enhance pedagogical precision, the prompt methodology can be improved by: (a) aligning dialogues with specific CEFR levels (e.g., B1 or C1) for comprehensible input (Krashen, 1985); (b) embedding concrete lexical and grammatical objectives through domain-specific terminology or syntactic structures; (c) enhancing pragmatic depth by including discourse markers for conversational fluency; and (d) enforcing coherent structure through a logical sequence (opening, core task, resolution). These specifications help move beyond generic dialogue generation toward tailored, objective-driven language learning resources.

A total of 4.500 dialogues were generated (1.500 from each model) to ensure a sufficient and balanced sample for comparative analysis. The same prompt was used for all models to create a controlled environment in which the model itself was the main variable. The generation process was standardized with a temperature of 0.7 to balance creativity and coherence. All outputs were collected and documented with their source model and parameters, creating the final analysis-ready corpus and ensuring transparency and reproducibility.

Representative samples from the dialogue corpus are provided below to illustrate the baseline performance of the selected models. Tables 1 through 3 display the raw, unprocessed outputs generated by each LLM using the standardized prompt. For each model, one representative dialogue is presented in its original Greek text.[3] These selections demonstrate the inherent variance in dialogue structure and linguistic nuance across the three models without any post-generation editing.

KriKri	**KriKri**
User: Θέλω να μάθω για τα είδη δανείων που προσφέρετε. **Agent:** Έχουμε 3 βασικές κατηγορίες: • Στεγαστικά δάνεια • Καταναλωτικά δάνεια • Επιχειρηματικά δάνεια **User:** Και ποια τα κριτήρια έγκρισης; **Agent:** Εισόδημα, πιστοληπτική ικανότητα, εξασφαλίσεις. **User:** Ευχαριστώ, θα ξανακαλέσω	**User:** Thelo na matho gia ta eidi daneion pou prosferete. **Agent:** Echoume 3 vasikes katigories: • Stegastika daneia • Katanalotika daneia • Epicheirimatika daneia **User:** Kai poia ta kritiria egkrisis; **Agent:** Eisodima, pistoliptiki ikanotita, exasfaliseis. **User:** Efcharisto, tha xanakaleso

Table 1. Raw Dialogue Output – Krikri (Beta)

Gemini	**Gemini**
User: Καλημέρα, ήθελα να ρωτήσω για ένα στεγαστικό δάνειο, αλλά είμαι πολύ μπερδεμένη με τα επιτόκια. **Agent:** Καλημέρα σας. Είναι λογικό, υπάρχουν πολλές επιλογές. Τι ακριβώς σας δυσκολεύει; **User:** Δεν ξέρω αν πρέπει να διαλέξω σταθερό ή μεταβλητό. Φοβάμαι τις αυξήσεις. **Agent:** Το σταθερό σας προσφέρει σιγουριά για ένα διάστημα, π.χ. 10 χρόνια, με την ίδια δόση. **User:** Και μετά τι γίνεται; Αν το επιτόκιο τότε είναι πολύ ψηλό; **Agent:** Μετά το δάνειο γυρίζει σε μεταβλητό, αλλά μπορείτε να κάνετε νέα συμφωνία τότε. **User:** Μάλιστα. Και τι ποσό μπορώ να πάρω; Το σπίτι κάνει 180.000 ευρώ. **Agent:** Εξαρτάται από το καθαρό οικογενειακό σας εισόδημα. Πόσο είναι περίπου; **User:** Μαζί με τον σύζυγό μου βγάζουμε 2.800 ευρώ το μήνα. **Agent:** Με αυτό το εισόδημα μπορείτε να πάρετε έως και 140.000 ευρώ δάνειο. **User:** Μα το σπίτι κάνει 180.000. Τα υπόλοιπα 40.000 πρέπει να τα έχω εγώ; **Agent:** Ακριβώς. Η τράπεζα χρηματοδοτεί έως το 80% της αξίας του ακινήτου. **User:** Υπάρχουν και άλλα έξοδα; Γιατί άκουσα για δικηγόρους και μηχανικούς. **Agent:** Ναι, υπολογίστε περίπου ένα 5% επιπλέον για φόρους, συμβολαιογράφο και τεχνικό έλεγχο. **User:** Πω πω, μαζεύονται πολλά. Ο μηχανικός τι ακριβώς κάνει; **Agent:** Ελέγχει τη νομιμότητα του σπιτιού και κάνει την εκτίμηση της αξίας για την τράπεζα. **User:** Αν η εκτίμηση βγει λιγότερο από 180.000; **Agent:** Τότε το δάνειο θα υπολογιστεί στο 80% της χαμηλότερης τιμής. **User:** Κατάλαβα. Είναι πολύ αγχωτικό. Μπορώ να έχω μια γραπτή προσφορά; **Agent:** Φυσικά. Θα χρειαστώ τα εκκαθαριστικά σας και τις ταυτότητες για να βγάλω μια προέγκριση. **User:** Θα σας τα φέρω αύριο. Πρέπει να κλείσω ραντεβού; **Agent:** Καλύτερα ναι, για να μην περιμένετε. Τι ώρα σας βολεύει; **User:** Στις 11 το πρωί είναι καλά; **Agent:** Σας έκλεισα για τις 11 με τον κύριο Παπαδάκη. Θα σας περιμένει. **User:** Σας ευχαριστώ πολύ, με βοηθήσατε να ξεκαθαρίσω τα πράγματα. **Agent:** Χαρά μου. Θα τα πούμε από κοντά αύριο.	**User:** Kalimera, ithela na rotiso gia ena stegastiko daneio, alla eimai poly berdemeni me ta epitokia. **Agent:** Kalimera sas. Einai logiko, yparchoun polles epiloges. Ti akrivos sas dyskolevei; **User:** Den xero an prepei na dialexo stathero i metavlito. Fovamai tis afxiseis. **Agent:** To stathero sas prosferei sigouria gia ena diastima, p.ch. 10 chronia, me tin idia dosi. **User:** Kai meta ti ginetai; An to epitokio tote einai poly psilo; **Agent:** Meta to daneio gyrizei se metavlito, alla boreite na kanete nea symfonia tote. **User:** Malista. Kai ti poso boro na paro; To spiti kanei 180.000 evro. **Agent:** Exartatai apo to katharo oikogeneiako sas eisodima. Poso einai peripou; **User:** Mazi me ton syzygo mou vgazoume 2.800 evro to mina. **Agent:** Me afto to eisodima boreite na parete eos kai 140.000 evro daneio. **User:** Ma to spiti kanei 180.000. Ta ypoloipa 40.000 prepei na ta echo ego; **Agent:** Akrivos. I trapeza chrimatodotei eos to 80% tis axias tou akinitou. **User:** Yparchoun kai alla exoda; Giati akousa gia dikigorous kai michanikous. **Agent:** Nai, ypologiste peripou ena 5% epipleon gia forous, symvolaiografo kai techniko elencho. **User:** Po po, mazevontai polla. O michanikos ti akrivos kanei; **Agent:** Elenchei ti nomimotita tou spitiou kai kanei tin ektimisi tis axias gia tin trapeza. **User:** An i ektimisi vgei ligotero apo 180.000; **Agent:** Tote to daneio tha ypologistei sto 80% tis chamiloteris timis. **User:** Katalava. Einai poly anchotiko. Boro na echo mia grapti prosfora; **Agent:** Fysika. Tha chreiasto ta ekkatharistika sas kai tis taftotites gia na vgalo mia proegkrisi. **User:** Tha sas ta fero avrio. Prepei na kleiso rantevou; **Agent:** Kalytera nai, gia na min perimenete. Ti ora sas volevei; **User:** Stis 11 to proi einai kala; **Agent:** Sas ekleisa gia tis 11 me ton kyrio Papadaki. Tha sas perimenei. **User:** Sas efcharisto poly, me voithisate na xekathariso ta pragmata. **Agent:** Chara mou. Tha ta poume apo konta avrio.

Table 2. Raw Dialogue Output – Gemini 2.5

[3] Greek text was transliterated into Latin script according to ISO 843, a standard system for the transliteration and/or transcription of Greek characters, to facilitate computational analysis.

ChatGPT	ChatGPT
User: Καλησπέρα, θέλω να κάνω μεταφορά χρημάτων μέσω mobile app.	**User:** Kalispera, thelo na kano metafora chrimaton meso mobile app.
Agent: Καλησπέρα. Χρειάζεστε εγγραφή στο mobile app και αριθμό λογαριασμού παραλήπτη.	**Agent:** Kalispera. Chreiazeste engrafi sto mobile app kai arithmo logariasmou paralipti.
User: Υπάρχει όριο ανά συναλλαγή;	**User:** Yparchei orio ana synallagi;
Agent: 500€ ανά συναλλαγή.	**Agent:** 500€ ana synallagi.
User: Υπάρχει κόστος;	**User:** Yparchei kostos;
Agent: Όχι, για εντός τράπεζας.	**Agent:** Ochi, gia entos trapezas.
User: Μπορώ να ορίσω επαναλαμβανόμενες μεταφορές;	**User:** Boro na oriso epanalamvanomenes metafores;
Agent: Ναι, ημερήσιες ή εβδομαδιαίες.	**Agent:** Nai, imerisies i evdomadiaies.
User: Υπάρχει ειδοποίηση για εκτέλεση;	**User:** Yparchei eidopoiisi gia ektelesi;
Agent: Ναι, SMS ή email.	**Agent:** Nai, SMS i email.
User: Μπορώ να ακυρώσω πριν εκτελεστεί;	**User:** Boro na akyroso prin ektelestei;
Agent: Ναι.	**Agent:** Nai.
User: Υπάρχει ιστορικό συναλλαγών;	**User:** Yparchei istoriko synallagon;
Agent: Ναι.	**Agent:** Nai.
User: Μπορώ να μεταφέρω σε άλλη τράπεζα;	**User:** Boro na metafero se alli trapeza;
Agent: Ναι, υπάρχει κόστος 0,3%.	**Agent:** Nai, yparchei kostos 0,3%.
User: Πόσος χρόνος χρειάζεται;	**User:** Posos chronos chreiazetai;
Agent: 1–2 εργάσιμες ημέρες για άλλες τράπεζες.	**Agent:** 1–2 ergasimes imeres gia alles trapezes.
User: Μπορώ να ορίσω φίλτρα ειδοποιήσεων;	**User:** Boro na oriso filtra eidopoiiseon;
Agent: Ναι, για ποσά πάνω από συγκεκριμένο όριο.	**Agent:** Nai, gia posa pano apo sygkekrimeno orio.
User: Υπάρχει ημερήσιο όριο συνολικά;	**User:** Yparchei imerisio orio synolika;
Agent: 1.000€ για mobile app.	**Agent:** 1.000€ gia mobile app.
User: Μπορώ να χρησιμοποιήσω QR code;	**User:** Boro na chrisimopoiiso QR code;
Agent: Ναι.	**Agent:** Nai.
User: Υπάρχει ειδοποίηση για αποτυχημένες μεταφορές;	**User:** Yparchei eidopoiisi gia apotychimenes metafores;
Agent: Ναι.	**Agent:** Nai.
User: Μπορώ να δω ισοτιμία πριν τη μεταφορά;	**User:** Boro na do isotimia prin ti metafora;
Agent: Ναι.	**Agent:** Nai.
User: Ευχαριστώ πολύ.	**User:** Efcharisto poly.
Agent: Χαρά μας να σας εξυπηρετούμε!	**Agent:** Chara mas na sas exypiretoume!

Table 3. Raw Dialogue Output – GPT-4.1

2.3 Evaluation framework

The final phase of the study comprised the evaluation of the generated dialogues through a combination of qualitative and quantitative assessment. The dataset comprises 1.500 clauses per model, all manually annotated according to a predefined evaluation schema. To ensure a systematic and replicable analysis, an evaluation rubric was developed to assess the dialogues across five principal dimensions:

- *Thematic Appropriateness and Variation:* This assessed how effectively each dialogue aligned with banking-related intents and how comprehensively it explored the range of expected use cases within the banking domain.
- *Structural and Pragmatic Dimensions of Dialogue Resolution:* This evaluated how coherent the dialogue was, whether turns followed logically, and whether the conversation concluded successfully.
- *Morphosyntactic and Lexical Accuracy:* This assessed the grammatical and spelling correctness of each dialogue, focusing on errors and unnecessary English code-switching or loanwords in the Greek text.
- *Politeness, Empathy, and Naturalness:* This dimension evaluated how human-like and socially appropriate the dialogues were. It captures whether the model uses polite phrasing, demonstrates empathy, or adapts responses naturally to the user's intent and context.

For the quantitative linguistic analysis (e.g., tokens, types, and frequency patterns), the corpus was processed using AntConc and Python. The evaluation of syntax, semantics, and naturalness, by contrast, was conducted through manual qualitative analysis. All clauses were evaluated using the same criteria to ensure consistency. Annotations were performed by a single evaluator; while this represents a limitation in terms of inter-annotator reliability, the systematic application of the

predefined schema across all models helped maintain consistency and transparency across the dataset.[4]

This multi-dimensional approach, combining quantitative and qualitative analysis, enabled a deeper understanding of the linguistic and pedagogical value of AI-generated dialogues, providing evidence of their potential for classroom use and further linguistic research.

3. Data Analysis

This section presents the analysis of the data collected for the study, focusing on the linguistic and communicative features identified within the generated corpus. The analysis aims to evaluate the structural integrity, lexical density, and pragmatic coherence of the synthetic dialogues. By employing a multi-dimensional approach, this section identifies the underlying patterns that characterize the output of each LLM, moving from aggregate quantitative metrics to granular qualitative assessments. The following subsections outline the analytical framework and present the key findings that inform the subsequent discussion and conclusions.

3.1 Linguistic Complexity and Density

A comparative quantitative analysis was conducted on the dialogues generated by the three LLMs based on their aggregate lexical and structural statistics. The dataset comprised total types and token counts, average dialogue steps, as well as total dialogue steps per model. To enable normalized cross-model comparison, three derived linguistic ratios were computed: the type-to-token ratio (TTR) to assess lexical diversity, tokens per dialogue steps to estimate verbosity per dialogue turn, and types per dialogue steps to approximate lexical richness (Table 4).

Model	Lemmas	Tokens	Av. Steps	Lemma/Token	Tokens/Step	Lemmas/Step	Tokens/Lemma
Krikri	10,338	79,221	4	0.13	19,805.25	2,584.5	7.66
Gemini	19,218	355,902	19	0.054	18,731.68	1,011.47	18.52
ChatGPT	13,194	20,959	18	0.63	1,164.39	733.00	1.59

Table 4. Tokenization and Lemmatization Statistics per Model

ChatGPT produced 13194 types and 202959 tokens across an average of 19 dialogue steps, yielding a TTR of 0.065 and the lowest tokens-per-dialogue-steps value of 7.11, indicating a less lexically diverse but highly compact generative pattern. Gemini generated 19218 types and 355902 tokens over 19 dialogue steps, resulting in the lowest TTR of 0.054 and the highest tokens-per-dialogue-steps ratio of 12.16, suggesting the least lexically diverse and most verbose output structure among the three models. Krikri, with 10338 types and 79221 tokens distributed across an average of only 6 steps, exhibited the highest TTR of 0.13 and the highest types-per-dialogue-steps ratio of 1.12. However, these results must be interpreted with caution, as the task specifications required a minimum of 8 dialogue steps per dialogue. Krikri's systematic non-compliance with this requirement naturally inflates its TTR and types-per-dialogue-steps ratio, since shorter dialogues provide less opportunity for lexical repetition. Therefore, Krikri's apparent lexical diversity may be an artifact of its shorter

[4] While the corpus consists of LLM-generated dialogues rather than authentic banking data, the author's prior experience with Omilia's real banking dialogue datasets provided a grounded understanding of expected dialogue structures, communicative patterns, and domain conventions. This background informs the evaluative criteria and allows for occasional comparisons with real-world dialogue output.

dialogue length rather than genuinely richer vocabulary usage. Overall, ChatGPT prioritizes compactness, Gemini produces extensive output with greater redundancy, and Krikri, despite appearing lexically diverse, falls short of the minimum dialogue step requirement, compromising its metrics comparability. Adherence to task specifications emerges as a crucial factor in ensuring valid cross-model comparisons.

The subsequent stage of this examination involves the analysis of the clausal distribution produced per dialogue, focusing specifically on syntactic depth and the interactional balance maintained between the Agent and the User (Table 5 and 6).

Model	Max Clauses (Agent|User)	Min Clauses (Agent|User)	Avg.Agent Clauses/ Dialogues	Avg.User Clauses/ Dialogues	Total Sum (Agent|User)
Krikri	14 | 11	2 | 1	3.20	2.46	4800|3696
Gemini	38 | 46	7 | 6	14.85	14.59	22280|21891
ChatGPT	21 | 17	2 | 2	8.02	6.81	12041|10219

Table 5. Comparative Clause Analysis of Agent–User Exchanges

Model	Avg. Clauses/ Dialogues	Aggregate Output
Krikri	5.67	8496
Gemini	29.45	44171
ChatGPT	14.84	22260

Table 6. Comparative Clause Analysis per dialogue

Regarding the syntactic dimension, the investigated models displayed markedly divergent profiles in structural complexity and interactional density (Table 5 and 6). ChatGPT generated an aggregate of 22260 clauses, with the Agent and User averaging 8.02 and 6.81 clauses per dialogue, respectively. This performance suggests a balanced and controlled syntactic depth, where the model maintains substantial structural complexity while successfully eliciting the high engagement necessary to sustain the dialogue toward the requested length. In contrast, Gemini produced a significantly more expansive output of 44171 aggregate clauses. This corpus is characterized by a markedly higher mean of 14.85 clauses for the Agent and 14.59 for the User. Such data indicates a “high extension” profile; while the model adheres to the requirement for a lengthy interaction, it achieves this through highly subordinate and multi-clausal structures that suggest a tendency toward extreme verbosity. Conversely, Krikri yielded a total of only 8496 clauses, with the Agent averaging 3.20 clauses and the User 2.46. This suggests a much lower level of syntactic complexity and shorter interaction style, indicating the model struggled to produce the structural complexity required to meet the prompt's main objective.

The minimum and maximum clause counts (as presented in Table 5) further support the trends identified in the previous analysis. Gemini demonstrates the highest level of structural complexity, with maximum clause counts reaching 38 for the Agent and 46 for the User, indicating its capacity to sustain longer and more elaborate dialogues. ChatGPT shows more moderate values (Agent: 21 | User: 17), reflecting a more concise and structurally efficient interaction style. In contrast, Krikri presents the lowest clause ceilings (Agent: 14 | User: 11) and minimal clause counts as low as one clause for the user, suggesting difficulty in supporting complex conversational exchanges. Overall, Gemini’s longer and more complex responses make it better suited for detailed storytelling and engaging dialogue than the other models.

3.2 Thematic Appropriateness and Variation

The following analysis examines how well the dialogue aligns with the intended banking domain and the range of expected user intents within that domain. It evaluates whether interactions remain contextually relevant to banking scenarios (e.g., transactions, account management, financial advice) and whether the dialogue demonstrates appropriate thematic coverage and variation across different banking use cases, without drifting into unrelated topics (Table 7).

Categories	Krikri	Gemini	ChatGPT	Focus
Transaction	18%	10%	40%	Functional & Operational Execution
Card	32%	25%	25%	Crisis Resolution & Risk Mitigation
Account Management & Onboarding	5%	10%	—	Lifecycle & Institutional Compliance
Fees, Charges & Claims	3%	—	—	Conflict Resolution & Grievance
Financial Solutions	24%	20%	15%	Capital Growth & Lending Strategy
Investment & Insurance	15%	—	12%	Wealth Management & Protection
Generic Questions & Loyalty	3%	—	—	Information Retrieval & Retention
Digital Support & Infra.	—	—	8%	Technical Access & Digital Ecosystem
Technical Support & Reliability	—	35%	—	System Uptime & Interface Integrity

Table 7. Distribution of Thematic Categories Across Model

The three datasets reveal distinct thematic priorities across models, reflecting a spectrum from traditional banking engagement to operational execution and technical troubleshooting. Krikri offers the broadest coverage across all nine categories, emphasizing Card (32%) and Financial Solutions (24%), which together comprise 56% of its volume, reflecting a user base oriented toward financial product engagement and lending. Krikri uniquely covers categories absent from other models, such as Fees, Charges & Claims (3%) and Generic Questions & Loyalty (3%), indicating a more diverse banking interaction profile. ChatGPT identifies a user base primarily focused on high-frequency operational execution, with Transactions accounting for 40% of total volume, which is the highest single-category concentration across all models. It also allocates 25% to Card-related interactions and is the only model alongside Krikri to cover Investment & Insurance (12%), while uniquely addressing Digital Support & Infrastructure (8%). However, Gemini introduces a notably different thematic profile, where Technical Support & Reliability (35%) emerges as the dominant category; this theme is entirely absent from both Krikri and ChatGPT. This, combined with Card (25%) and Financial Solutions (20%), accounts for 80% of Gemini's total volume, suggesting a narrower but more technically focused interaction model.

The consistently high Card interaction volume across all three models (25–32%) indicates that crisis resolution and risk mitigation remain universal concerns in digital banking. Overall, these patterns suggest modern banking user experience, as modeled by LLMs, encompasses not only traditional financial information-seeking but increasingly prioritizes system reliability and operational efficiency as core dimensions of customer relationship.

3.3 Structural and Pragmatic Dimensions of Dialogue Resolution

In the analysis of customer service interactions, the resolution phase serves as a critical indicator of interaction completeness and participant satisfaction. To systematically evaluate the termination dynamics within the examined dataset, each dialogue was classified into one of three distinct closure categories based on the final speaker and the semantic context of the concluding utterance.

The “Agent Closed” category designates dialogues that achieve a natural and formal termination with the customer service representative delivering the final utterance. In these instances, the agent characteristically confirms the successful execution of a requested operational task or employs standard closing formalities to gracefully conclude the exchange. For example, an interaction is classified under this label when the agent validates the completion of a process, such as stating “Egine. Se 2 lepta tha einai etoimi gia chrisi. (Done. In 2 minutes, it will be ready for use)”. Alternatively, the classification applies when the agent outlines the subsequent procedural steps or offers a formal valediction, evidenced by phrases like “Fysika, tha katevasoume tin efarmogi mazi sto katastima. Efcharistoume poly pou epikoinonisate mazi mas. (Of course, we will download the app together at the branch. Thank you very much for contacting us.)”.

The “User Closed” category characterizes interactions that conclude with the customer delivering the final message, signaling that their primary inquiry has been sufficiently addressed. This classification occurs when the user explicitly acknowledges the provided solution, articulates gratitude, or offers a concluding salutation, thereby rendering further intervention from the agent unnecessary. Representative linguistic markers for this category include affirmations of proposed solutions or next steps, such as “Teleia, to dokimazo tora. (Perfect, I am trying it now.)”, or expressions of satisfaction regarding procedural timelines, such as "Oraia, elpizo na prolavo ta eisitiria. (Great, I hope I catch the tickets in time.)”. Furthermore, straightforward expressions of appreciation that naturally terminate the dialogue, including “Teleia, efcharisto gia tin grigori apantisi. (Perfect, thank you for the quick response)”, are unequivocally assigned to this label.

Finally, the “Pending / Dropped” category identifies dialogues that terminate abruptly following a user utterance, leaving the interaction demonstrably unresolved. These instances are characterized by a final customer message that conveys a direct inquiry or an explicit directive, which structurally mandates a subsequent response, confirmation, or action from the agent that is absent from the transcript. Therefore, the dialogue is considered incomplete. Illustrative examples from the corpus include unresolved directives, such as “Efcharisto, kante to tora. (Thank you, do it now)”, where the agent's confirmation of execution is missing. Similarly, the category encompasses terminal interrogatives necessitating further procedural guidance, exemplified by “Teleia, ti chartia na sas fero? (Perfect, what papers should I bring you?)”. This category is particularly relevant for models that produce shorter dialogues, as insufficient dialogue length increases the likelihood of premature termination before communicative closure is achieved.

To assess the pragmatic efficacy of the investigated models, the following table (Table 8) provides a quantitative basis for evaluating dialogue closure and the successful resolution of communicative intent.

Categories	Krikri	ChatGPT	Gemini
Agent Closed	29%	67%	57%
User Closed	13%	29%	36%
Pending/Dropped	58%	4%	7%

Table 8. Intent Resolution and Dialogue Closure Across Models

ChatGPT demonstrates the most efficient, goal-oriented structure with a 67% agent-led closure rate and the lowest Pending/Dropped rate (4%), effectively bridging literal input and intended outcome for successful task completion. Gemini shows a similarly robust profile (57% agentled closures, 7% unresolved interactions) while achieving the highest user-led closure rate (36%), fostering greater interactional symmetry where users feel empowered to conclude dialogues independently. Conversely, Krikri exhibits a markedly different pattern, with 58% of dialogues classified as Pending/Dropped. Its average dialogue length of 6 turns falls below the required minimum of 8, contributing to failure in achieving proper communicative closure, with only 29% agent-led and 13% user-led closures, indicating a significant deficit in illocutionary fulfillment. Overall, ChatGPT remains the superior model for goal-directed discourse, Gemini offers a balanced alternative with strong resolution rates and greater user agency, while Krikri's high number of unresolved interactions highlights the critical relationship between dialogue completeness and effectiveness.

3.4 Morphosyntactic and Lexical Accuracy

This section analyzes the morphosyntactic and lexical accuracy of the three LLMs, comparing the frequency and types of errors and examining how English influences their Greek dialogues. The analysis of morphosyntactic accuracy across the three models revealed significant variation in grammatical performance.

Krikri exhibited the highest frequency of morphosyntactic errors, with at least one grammatical mistake identified in approximately 35% of its dialogues. The errors included syntactic redundancy (e.g., “Tha sas perimenoume tin karta sas. (e.g. unnecessary repetition of the possessive pronoun 'We will be waiting [you] for your card. [y]’)”), incorrect verb inflection (“I karta mou klepike” instead of “eklapi(stolen)”), and gender/number disagreement in article–noun combinations (“mia minyma(one message)”). In addition, word order and pronoun placement errors such as “Tha perimenoume sas gia tis ypografes. (Will wait you for the signatures.)” indicate limited control over Greek clitic usage and argument structure. This high error rate aligns with findings from Section 3.1, where Krikri's average dialogue length fell below the required minimum of 8, and its elevated Pending/Dropped rate of 58% (Section 3.3) suggests the model's output quality is compromised by structural incompleteness.

ChatGPT demonstrated a notably higher level of grammatical accuracy, with only 15% of dialogues containing at least one morphosyntactic deviation. The most frequent issues concerned article omission and preposition omission (e.g., “anoigma logariasmou gia douleia (opening a bank account for work)” instead of “(...gia tin douleia (for the work))”), case errors (e.g., “Nai, gia misthos” instead of “Nai, gia mistho (Yes, for salary)”), and elliptical or non-standard clause structure (“Prepi rantevu; (Need appointment?)” instead of “Chreiazetai na kleiso rantevu; (Do I need to book an appointment?)”). These patterns suggest a partial command of Greek grammatical morphology, particularly regarding declension and syntactic completeness.

Gemini produced dialogues that were entirely free of grammatical errors. All utterances adhered to standard Greek morphosyntax, demonstrating native-like control of agreement, case marking, and article use. While this finding is notable, it should be interpreted in the context of Gemini's output profile: as established in Section 3.1, Gemini produced the highest token count (355902) with the lowest TTR (0.054), suggesting that its grammatical accuracy may partly stem from a more repetitive and formulaic output structure that reduces the likelihood of morphosyntactic deviation.

Overall, the results suggest a clear gradation of grammatical competence, with Gemini achieving full morphosyntactic accuracy, ChatGPT displaying moderate reliability, and Krikri revealing systematic weaknesses, particularly in verb morphology and pronoun syntax.

From the analysis of the data, two additional observations emerged that illustrate the influence of English on Greek in model-generated dialogues. The first concerns words and phrases translated directly from English into Greek, where the models tend to reproduce English syntactic or idiomatic structures without proper adaptation to Greek usage. For example, the English expression “pay attention” is sometimes rendered literally as “plirono prosochi” instead of the natural Greek equivalent “dino prosochi (give attention)”. This leads to instances of translated Greek, where the text is grammatically correct but stylistically unnatural. The second case involves the direct use of English words and expressions without translation (e.g. “ATM”, “3D Secure”), a phenomenon reflecting both real-world linguistic borrowing and the dominance of English in digital and professional communication. These two observations highlight the extent to which large language models are shaped by English-dominant data sources and demonstrate how this influence manifests in generated Greek dialogues.

From the first category, only Krikri and ChatGPT contain examples of literal translations from English, such as “keep your word → kratas ti lexi sou (natural: kratas ton logo sou)”,” end to end → akri me akri (natural: apo akri se akri)”, and “make sense → kano noima (natural: vgazo noima)”. Gemini, on the other hand, does not produce such literal translations. In contrast, all models show a high percentage of English loanwords, reflecting the widespread incorporation of untranslated English terms in Greek dialogues. Based on our analysis, the categories of these occurrences can be summarized as follows:

• Direct Adoption: Instead of using Greek terms like ilektroniki trapeziki, the text consistently uses “e-banking”. Other examples include “app”, “site”, “password”, and “link”.
• Acronyms without Greek translation: Technical banking terms like IBAN, SWIFT, SEPA, and ESG are used in their original English forms.
• Banking Specific Terminology: Terms such as “P/E ratio” and “ED Secure” appear directly in English, often requiring the agent to explain them immediately after use.
• Common Loanwords: Words like “okay” and “video” (via Teams/Zoom) are fully integrated into the dialogue.

3.5 Politeness, Empathy, and Naturalness

This section assesses the pragmatic and stylistic quality of the dialogues, focusing on polite markers (e.g., “please”, “thanks”, “sorry for the inconvenience”), empathetic responses that acknowledge user emotions or frustration, and natural phrasing that reads like authentic human conversation rather than rigid machine output.

All three models consistently employ the plural of politeness to maintain professional distance. In Krikri, agents greet users with “Kalimera sas! Pos boro na voithiso; (Good morning! How can I help?)” and use collaborative softening like “As to doume mazi (Let's see it together)”. ChatGPT maintains formal address (“Kalispera sas. Boreite na mou dosete... (Good evening. Can you give me...?)”), while Gemini adds honorifics such as “kyria Maria (Ms Maria)”, personalizing responses while preserving professional tone. Across all models, phatic expressions including greetings, supportive closings (“Kali synecheia (Have a good rest of your day)"), and acknowledgement fillers (“Katalavaino (I understand)”, “Malista (I see)”) maintain conversational flow and social rapport.

However, there are clear differences in empathy. Krikri demonstrates foundational empathy through standardized phrases such as “Katalavaino tin anisychia sas (I understand your concern)”. Yet, given that 58% of its dialogues were classified as Pending/Dropped (Section 3.3) and its average dialogue length fell below the required minimum of 8 steps (Section 3.1), opportunities for sustained

empathetic engagement are structurally limited, as most dialogues terminate before full rapport can develop. ChatGPT provides more structured empathetic sequences, incorporating markers such as “Mi stenochoriesie (Do not worry)” and “Den chreiazetai anchos (There is no need for stress)", guiding users from anxiety ("Eimai ligo anchomenos (I am a bit nervous)”) to relief (“Oraia, nai, egine! (Great, yes, it's done!)”). Its 67% agent-led closure rate (Section 3.3) ensures that these empathetic arcs consistently reach full resolution. Gemini delivers the most human-like interactions, synthesizing formal politeness with dynamic phatic expressions and supportive reassurance such as “Min anisycheite (Don't worry)” and “Eimai edo gia na sas voithiso (I am here to help you)”, even when responding to high-stress prompts like “Den echo alla lefta pano mou! (I don't have any more money on me!)”. Furthermore, Gemini maintains a superior conversational flow through the adept use of fillers - such as “Malista (Certainly/I see)” and “Katalavaino (I understand)” - and seamless turn-taking, which leads the user to a state of relief, often expressed as “A, afto voithaei poly (Ah, that helps a lot)”. With only 7% Pending/Dropped dialogues and the highest user-led closure rate of 36% (Section 3.3), Gemini's empathetic engagement is supported by strong structural completeness, allowing for fully developed conversational arcs.

Overall, Gemini exhibits the most balanced and emotionally responsive dialogue, supported by natural phrasing and strong completion rates. ChatGPT follows closely, combining structured empathy with the highest resolution rate. Krikri, despite displaying foundational politeness, is constrained by its structural shortcomings.

4. Proposed Pedagogical Application

The findings of this study carry relevant implications for language pedagogy in both first (L1) and second (L2) language education. The proposed instructional framework shifts away from a traditional binary of "correct" versus "incorrect" texts, instead utilizing a multidimensionally annotated corpus. Rather than diverging from established approaches, this method aligns with data-driven learning (DDL) methodologies (Boulton & Cobb, 2017; Crosthwaite & Baisa, 2023), drawing on a multidimensionally annotated corpus of LLM-generated dialogues. Crosthwaite and Baisa (2023) argue that the integration of generative AI with DDL represents a logical and mutually beneficial connection, offering several advantages, including fewer barriers compared to traditional corpus query tools.

In this framework, curated dialogues are not defined as manually sanitized or pre-corrected data, but as LLM-generated dialogues that have been preserved in their original state and layered with granular metadata and annotations. This annotation-led approach transforms the dataset into a flexible pedagogical laboratory where the educator acts as the primary curator, leveraging the metadata to filter and select specific data subsets that align with distinct instructional objectives.

Through the deliberate selection of dialogues characterized by reduced accuracy or naturalness, educators may leverage non-curated AI-generated content as pedagogical instruments for diagnostic analysis and reflective critique. Based on the findings of this study, Krikri's output, which exhibited the highest morphosyntactic error rate (approximately 35% of dialogues, Section 3.4) and frequent structural incompleteness (58% Pending/Dropped, Section 3.3), provides particularly rich material for such activities.

• Diagnostic error-spotting: Educators can filter dialogues based on morphosyntactic and lexical accuracy to design targeted error-spotting tasks. Learners are challenged to identify inflectional errors, inappropriate English loanwords, or register inconsistencies. Crucially, they must justify why a machine-generated utterance feels unnatural in a professional banking context, moving beyond passive observation to active analysis (James, 2013).

• Metalinguistic and Editorial Analysis: Native speakers can treat low-scoring outputs as "flawed drafts". By analyzing how the machine mismanages the subtle socio-pragmatic nuances of Greek professional discourse - such as the literal translations from English identified in Krikri and ChatGPT (e.g., "plirono prosochi", Section 3.4) - L1 students develop the robust metalinguistic awareness required for high-level editorial and professional communication.

Conversely, dialogues filtered for high structural resolution and thematic appropriateness serve as authentic models for production-oriented and communicative activities. The findings suggest that Gemini and ChatGPT, which demonstrated strong dialogue completion rates and superior pragmatic quality (Section 3.5), provide the most suitable material for these tasks.

• Task-Based Language Teaching (TBLT): Within a TBLT framework (Ellis, 2003), high-scoring curated dialogues move learners from mere imitation toward meaningful problem-solving (Salies, 1995; Long, 2015; Bahriyeva, 2021). For example, ChatGPT's Transaction-heavy dialogues (40%, Section 3.2) offer structured scenarios for practicing operational banking tasks, while Gemini's Technical Support dialogues (35%, Section 3.2) expose learners to troubleshooting vocabulary and discourse patterns.
• Roleplay and Specialized Vocabulary Development: Learners can adapt and perform role-plays, such as negotiating a loan or resolving a service issue, using the AI dialogues as foundational scripts (Fu & Li, 2025). This approach is strengthened by integrating specialized pedagogical lexicons (Tarp, 2008), which are custom tools tailored with register annotations, semantic nuances, and domain-specific tags designed for educational purposes. When used alongside the curated dialogues, these lexicons help learners cross-reference unfamiliar financial terminology with authentic usage patterns, applying complex vocabulary within coherent discourse contexts.

LLMs are typically trained on broad, multinational datasets, often resulting in outputs that are linguistically correct but culturally misaligned, such as being overly formal, too indirect, or heavily influenced by Anglocentric norms for the Greek service context. The annotated metadata facilitates deep exploration of these intercultural dimensions. By filtering for specific pragmatic scores, educators can initiate comparative activities where students contrast an authentic, human-generated Greek request with its culturally misaligned LLM counterpart. For instance, the English influence patterns identified in Section 3.4 (literal translations and untranslated English terms) provide a tangible basis for discussing intercultural variations in politeness, directness, and linguistic borrowing. Consequently, learners enhance their awareness of how language encodes cultural expectations, enabling them to make informed pragmatic choices in cross-cultural professional communication.

Ultimately, this pedagogical framework repositions AI-generated dialogues as more than mere text generators; they become active catalysts for critical reflection. By bridging computational output with human interpretation, this approach allows learners to deeply engage with linguistic precision, pragmatic appropriateness, and intercultural competence within specialized domains like banking.

5. Future work

Building on the finding that models exhibit distinct linguistic profiles future research will focus on expanding and refining the corpus to enhance both its linguistic scope and pedagogical applicability. One key direction involves increasing the diversity of banking scenarios to include emerging financial technologies such as digital banking and AI-driven customer support. Incorporating these new domains will enable the analysis of evolving communicative practices and provide richer material for language instruction in contemporary financial contexts.

A crucial step in validating the corpus will be its systematic comparison with real customer–service interactions collected from authentic banking data (e.g., anonymized transcripts or simulated training

datasets). Such comparisons will help assess the linguistic naturalness, pragmatic accuracy, and contextual realism of the LLM-generated dialogues. Insights from this comparison will guide future refinements in both corpus construction and prompt engineering, ensuring closer alignment with real communicative practices in the banking sector.

The pedagogical potential of the corpus will be further examined through empirical testing and classroom-based evaluation with language learners and instructors. This phase will involve controlled studies assessing learners' engagement, comprehension, and communicative competence when exposed to LLM-generated dialogues compared to authentic materials. Additionally, experimental designs may incorporate pre- and post-intervention assessments to measure linguistic gains and task performance, while teacher feedback and learner perceptions will be collected to evaluate usability and pedagogical effectiveness. Findings from these empirical investigations will not only validate the corpus as a teaching resource but also inform evidence-based guidelines for integrating LLM-generated materials into domain-specific language education.

Through these developments, future work aims to establish a scalable, empirically validated, and pedagogically effective framework for the use of large language models in domain-specific language education.

References

Bahriyeva, N. (2021). Teaching a language through role-play. *Linguistics and Culture Review*, *5*(S1), 1582-1587 https://doi.org/10.21744/lingcure.v5nS1.1745

Baskara, R., & Mukarto, M. (2023). Exploring the implications of ChatGPT for language learning in higher education. *Indonesian Journal of English Language Teaching and Applied Linguistics, 7*(2), 343-358Boulton, A., & Cobb, T. (2017). Corpus Use in Language Learning: A Meta-Analysis. *Language Learning, 67*, 348-393. https://doi.org/10.1111/lang.12224

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). *Language Models are Few-Shot Learners.* Advances in Neural Information Processing Systems, 33, 1877-1901. https://doi.org/10.48550/arXiv.2005.14165

Crosthwaite, P., & Baisa, V. (2023). Generative AI and the end of corpus-assisted data driven learning? Not so fast! *Applied Corpus Linguistics*, 3(3). https://doi.org/10.1016/j.acorp.2023.100066

Geroimenko, V. (2025). Key Principles of Good Prompt Design. In V. Geroimenko (Ed.), *The Essential Guide to Prompt Engineering*. Springer Nature. Switzerland. https://doi.org/10.1007/978-3-031-86206-9_2

Godwin-Jones, R. (2021). Big data and language learning: Opportunities and challenges. *Language Learning & Technology*, 25(1), 4–19. https://doi.org/10.64152/10125/44747

James, C. (2013). *Errors in language learning and use: Exploring error analysis*. Routledge.

Jeon, J., & Lee, S. (2023). Large language models in education: A focus on the complementary relationship between human teachers and ChatGPT. Education and Information Technologies, 28, 15873-15892.

Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for Language Teaching and Learning. *RELC Journal*, 54(2), 537-550. https://doi.org/10.1177/00336882231162868

Long, M. H. (2015). *Second language acquisition and task-based language teaching*. Wiley- Blackwell.

Ng, D. T. K., Leung, J. K. L., Su, J., Ng, R. C. W., & Chu, S. K. W. (2023). Teachers' AI digital competencies and twenty-first century skills in the post-pandemic world. *Educational Technology Research & Development*, 71, 137–161. https://doi.org/10.1007/s11423-023-10203-6

Seo, K., Tang, J., Roll, I., Fels, S., & Yoon, D. (2021). The impact of artificial intelligence on learner–instructor interaction in online learning. *International Journal of Educational Technology in Higher Education*, 18, 1–23. https://doi.org/10.1186/s41239-021-00292-9

Salies, T.G. (1995). *Teaching Language Realistically [microform]: Role Play Is the Thing*. [ERIC document No. ED424753]. ERIC. https://eric.ed.gov/?id=ED424753

Tarp, S. (2008). *Lexicography in the Borderland between Knowledge and Non-Knowledge. General Lexicographical Theory with Particular Focus on Learner's Lexicography*. (Lexicographica. Series Maior, Volume 134. Tübingen: Max Niemeyer Verlag https://doi.org/10.1093/ijl/ecp030

Fu, X., & Li, Q. (2025). Effectiveness of role-play method: A meta-analysis. *International Journal of Instruction*, 18(1), 309-324. https://doi.org/10.29333/iji.2025.18117a

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., & Zhou, D. (2022). *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.* Advances in Neural Information Processing Systems, 35, 24824-24837. https://doi.org/10.48550/arXiv.2201.11903

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., ... Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with ChatGPT. *arXiv Preprint*, 2302.11382. https://doi.org/10.48550/arXiv.2302.11382

Williamson, B., & Eynon, R. (2020). Historical threads, missing links, and future directions in AI in education. *Learning, Media and Technology*, 45, 223–235. https://doi.org/10.1080/17439884.2020.1798995

Alexandra Fiotaki (alexandra.fiotaki@gmail.com) is an Adjunct Lecturer at the University of Athens, Department of Philology, teaching "Language Technology and Education". She holds an interdisciplinary interuniversity Master in Language Technology "Technoglossia" (UOA, NTUA & ILSP) and a PhD in Computational Linguistics, fully funded by GSRT & HFRI. Her doctoral research is a semasio-syntactic study of the Sequence of Tense phenomenon in Greek and an implementation within a computational grammar for Greek. She participated in several research projects on corpus linguistics and natural language understanding, focusing on data processing, multilingual computational grammars, and the development of educational technologies and interactive digital platforms. She also works as an NLU Specialist and NLG Data Pipeline Engineer at Omilia within the Department of Research and Machine Learning. She works on large-scale projects for domain-agnostic entities and intents, and has extensive experience in prompt engineering and the design of task specific Large LLMs. His professional work includes the management, annotation, and preprocessing of raw text data for model training across sectors such as banking and telecommunications. Her research interests lie in computational linguistics, corpus linguistics, and educational technology, extending to semantics, syntax, morphology, AI-mediated language learning, and the use of language technologies in education.