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.

Acknowledgements

This work was partially supported by: Omilia - Conversational Intelligence and the ELOQUENCE project (grant number 101070558), funded by the European Union under the Horizon 2020 Program.

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.