INTELLIGENT SYSTEMS AND THE APPLICATION OF ARTIFICIAL INTELLIGENCE IN FOREIGN LANGUAGE LEARNING: AN EMPIRICAL STUDY BASED ON THE CHATGPT PLATFORM

Aleksandar Ćosić*
[0009-0002-1312-4013]

Faculty of Organizational Sciences,
University of Belgrade,
Belgrade, Srebia

Abstract:

This paper presents the results of a controlled experiment conducted with 120 participants, examining the effects of the ChatGPT platform (GPT-4o) on learning English as a foreign language compared to traditional instruction. The experiment lasted 10 weeks, and participants were divided into an experimental group (N=60) and a control group (N=60). Outcomes measured included speaking skills, writing accuracy and lexical competence. The experimental group achieved statistically significant improvements across all measured areas. The paper integrates a theoretical foundation from a review of contemporary literature on intelligent systems, adaptive learning, and the pedagogical effects of AI technologies in language education, with particular attention to the capabilities of the ChatGPT platform, ethical challenges, and future directions for development.

Keywords:

Chatgpt--4o , Artificial Intelligence, Foreign Language Learning, Adaptive Learning, LLM.

INTRODUCTION

The rapid development of artificial intelligence has transformed educational technology from passive tools into active participants in the process of knowledge acquisition. In the field of foreign language learning, this transformation is particularly significant, as establishing communicative competence requires a high degree of personalization, regular feedback, and authentic interaction all of which are elements that traditional instructional models struggle to provide adequately for every learner individually. The integration of AI technologies into foreign language instruction today represents a shift from a uniform to a dynamic, adaptive learning environment.

A particularly prominent role in this transformation belongs to ChatGPT, a conversational AI system developed by OpenAI, based on the GPT-4o architecture. Unlike dedicated educational platforms, ChatGPT is a widely accessible tool that learners can use without specialized technical infrastructure, making it exceptionally suitable for application in higher education.

Correspondence:
Aleksandar Ćosić

e-mail:
acosic77@gmail.com

DOI: 10.15308/Sinteza-2026-432-437

Its ability to conduct flexible, contextually adapted conversations, correct errors, explain grammatical rules, and generate authentic language examples makes it a strong candidate for use in foreign language instruction.

Research over the past decade has recorded a rapid surge of interest in this field within the academic community. A review of empirical research studies has shown that AI-supported technology yields promising results in improving language learning, particularly in areas such as writing quality, assessment accuracy, and learner engagement. At the same time, systematic reviews identify as most effective those systems incorporating natural language processing, speech recognition, and automated feedback, especially when aligned with constructivist, communicative, or task-based approaches.

The aims of this paper are: (1) to describe the methodology of applying the ChatGPT platform in teaching English as a foreign language, (2) to present the methodology and results of the experiment, (3) to analyze the pedagogical effects in light of the relevant literature, and (4) to examine ethical challenges and future directions for development.

2. THEORETICAL BACKGROUND AND LITERATURE REVIEW

2.1. CONVERSATIONAL AI AND LARGE LANGUAGE MODELS IN EDUCATION

ChatGPT, as a representative of generative large language models (LLMs), functions as a general-purpose conversational platform capable of assuming various pedagogical roles: tutor, interlocutor, writing corrector, or exercise generator. Unlike traditional intelligent tutoring systems (ITS) built with fixed pedagogical logic, ChatGPT demonstrates high flexibility in adapting to learner levels and needs through natural dialogue.

The modern approach to applying LLMs in language instruction relies on prompt engineering, which teachers and learners use to direct interactions toward specific pedagogical goals: practising speech patterns, essay writing with correction, conversational practice in the target language, or grammar explanation. LLMs not only provide immediate and contextually relevant feedback, but also adapt their responses to a participant's prior interactions, creating a more personalized learning experience. University-level research confirms that students are generally satisfied with using ChatGPT for foreign language learning, though concerns around data privacy and the risk of misuse require ongoing pedagogical guidance.[1]

2.2. CHATBOT AGENTS AND CONVERSATIONAL SYSTEMS

AI-based chatbot agents have become one of the most popular tools for foreign language practice. A review of empirical studies published between 2020. and 2025. shows that chatbots are particularly effective in improving speaking fluency, writing quality, and learner motivation. Research has identified that the enhancement of AI-mediated speaking tasks increases learners' willingness to communicate, while studies have demonstrated advantages in vocabulary acquisition and grammar correction[2]. Of particular significance is the finding that chatbots reduce language anxiety, which is one of the key factors impeding the development of communicative competence in adult learners. ChatGPT stands out in this context due to its ability to sustain multiple, extended conversations in which it retains the context of prior exchanges within the same session, making it especially suited for simulating authentic language situations.

2.3. ADAPTIVE LEARNING AND AUTOMATED ASSESSMENT SYSTEMS

Adaptive systems use machine learning algorithms to personalize content. Although ChatGPT is not a classical adaptive system with an explicit learner model, its flexibility allows teachers to design structured sessions in which the level of difficulty is dynamically adjusted based on learner errors and responses. Automated writing evaluation (AWE) systems based on NLP algorithms, such as GPT-4o, enable the detection of grammatical errors, assessment of lexical richness, and provision of immediate feedback.

A meta-analysis of 46 empirical studies published between 2022 and 2025 demonstrated a statistically significant medium-to-large effect of AI on language learning ($g = 0.74$; 95% CI [0.57, 0.92]; $p < 0.001$) across all skills, with vocabulary showing the strongest effects. [3] This finding was instrumental in shaping the experimental protocol, which places particular emphasis on vocabulary development as the foundation for the growth of all other skills.

3. METHODOLOGY OF CHATGPT PLATFORM APPLICATION

3.1. PEDAGOGICAL PROTOCOL

For the purposes of this experiment, a structured pedagogical protocol was developed for the application of ChatGPT (GPT-4o, version available from January 2024), organized into three functional modules:

Conversational Practice Module - Participants engaged in guided conversations with ChatGPT using pre-defined prompts designed by the teacher. Example topics included everyday situations (shopping, travel, work), academic discourse, and debate topics. ChatGPT was instructed to respond at the B1/B2 level, correct errors at the end of each exchange, and explain corrections in context.

Writing and Correction Module - Participants wrote short essays and paragraphs, then submitted them to ChatGPT with a prompt requesting analytical correction (grammar, cohesion, lexical richness, pragmatic appropriateness). ChatGPT generated detailed feedback according to a pre-defined structured format included in the system prompt by the teacher.

Vocabulary and Grammar Module - The teacher created a bank of prompt templates for targeted practice of lexical and grammatical structures: generating examples, sentence completion, error identification in given texts, and mini-quizzes. Research on EFL learners' prompting behavior shows that students typically cycle through trial-and-error prompting strategies, and that explicit training in prompt construction significantly improves the quality of AI-generated responses and deepens engagement with writing tasks.[4]

3.2. PROMPT ENGINEERING AND INSTRUCTIONAL DESIGN

A central methodological contribution of this study is the development of a structured prompt engineering framework tailored specifically to the B1 CEFR level. Prompts were designed according to three operational parameters: (1) role assignment, in which ChatGPT was assigned a specific pedagogical persona (e.g., "You are a patient English tutor working with an intermediate-level adult learner"), (2) task specification, providing a clearly delimited instructional objective per session and (3) output format constraints, specifying the structure of expected responses (e.g., "First provide corrected text, then list each error with an explanation in no more than two sentences").

Each module operated with a distinct system prompt loaded at the beginning of each session. These system prompts were iteratively refined across the first two weeks of the experiment based on qualitative observations of response quality. The final prompt library consisted of 47 validated prompt templates distributed across the three modules. To ensure consistency, all participants in the experimental group used identical prompt templates for structured exercises, while open-ended conversational prompts were individualized under teacher guidance.

3.3. SESSION STRUCTURE AND WEEKLY SCHEDULE

Each of the three weekly ChatGPT sessions (45 minutes each) followed a standardized internal structure: (a) a 5-minute warm-up in which the participant re-established context with the AI using a continuation prompt from the prior session, (b) a 30-minute core task phase dedicated to the assigned module and (c) a 10-minute reflective closing phase in which participants asked ChatGPT to summarize the key linguistic points covered and to generate three personalized practice items for self-study. The weekly face-to-face session with the teacher was reserved for group debriefing, error pattern analysis across participants, and scaffolded discussion of language awareness raised during AI interactions.

Module rotation followed a weekly cycle: Week 1 - Conversational Practice (Mon/Wed) + Writing and Correction (Fri); Week 2 - Vocabulary and Grammar (Mon/Wed) + Conversational Practice (Fri); Week 3 - Writing and Correction (Mon/Wed) + Vocabulary and Grammar (Fri) with the cycle repeating across the 10-week intervention period. This design ensured that each module received equal instructional time while preventing habituation effects associated with repetitive task types.

3.4. ROLE OF THE TEACHER

The teacher retained a key role in: (a) creating and selecting pedagogical prompts, (b) monitoring the quality of ChatGPT responses, (c) facilitating reflective discussions on errors and learning strategies, and (d) assessing participant progress through formative evaluation.

Beyond prompt authorship, the teacher functioned as a critical mediator between the learner and the AI system. This supervisory function is consistent with recommendations from the literature on human-in-the-loop AI-enhanced instruction, which emphasizes that unmonitored AI interaction in language learning carries risks of reinforcing fossilized errors.

3.5. TECHNOLOGICAL INFRASTRUCTURE AND ACCESS CONDITIONS

All participants accessed ChatGPT (GPT-4o) via the free-tier web interface at chat.openai.com, without the use of API integrations or premium subscriptions. Sessions were conducted on participants' personal devices (laptop or desktop computers were recommended).

3.6. EXPERIMENTAL DESIGN AND SAMPLE

The experiment was conducted over 10 weeks (September-December 2025.) with a sample of 120 adult participants (mean age 24.3 ± 3.7 years), students enrolled in an English language course at the B1 level according to the Common European Framework of Reference for Languages (CEFR). Participants were randomly assigned to two groups:

Experimental group (N = 60): instruction supported by the ChatGPT platform (3 x 45 min per week with ChatGPT + 1 × 45 min with the teacher)

Control group (N = 60): exclusively traditional instruction (4 x 45 min per week with the teacher)

Randomization was carried out using stratified sampling by gender and pre-test scores to ensure group equivalence. All participants signed an informed consent form that included a description of the purpose of data collection and their right to withdraw.

Participants in the experimental group completed a 2 x 90-minute introductory training session on the effective use of ChatGPT for language learning, covering prompting techniques, critical evaluation of AI responses, and recognition of potential system errors.

3.7. MEASUREMENT INSTRUMENTS

The following outcomes were measured:

- Speaking skills - standardized IELTS Speaking format test
- Writing accuracy - essay task graded on holistic and analytic scales
- Lexical competence - Vocabulary Levels Test (VLT)

Measurement was conducted on a pre-test and post-test basis. Pre-tests were administered during the first week prior to any intervention, while post-tests were administered in the final week of the 10-week period. The interval between pre-tests and post-tests was therefore a minimum of nine instructional weeks for all participants, ensuring sufficient exposure to the respective instructional conditions before outcome assessment.

3.8. SPEAKING ASSESSMENT PROTOCOL

Speaking skills were assessed using a standardized IELTS Speaking format test administered individually by a trained examiner blind to group assignment. Each speaking test consisted of three parts mirroring the official IELTS format: Part 1 (introductory questions, 4–5 minutes), Part 2 (individual long turn with cue card, 3–4 minutes), and Part 3 (two-way discussion, 4–5 minutes). Responses were scored on the official IELTS band descriptors across four criteria: Fluency and Coherence, Lexical Resource, Grammatical Range and Accuracy, and Pronunciation. Final scores represent the mean of rater scores.

3.9. WRITING ASSESSMENT PROTOCOL

Writing accuracy was measured using a two-part essay task: a 150-word descriptive paragraph and a 250-word argumentative essay, both completed under controlled, timed conditions (60 minutes total) without access to external resources or AI tools. Essays were evaluated using both a holistic scale (aligned with IELTS Writing Task 2 band descriptors) and an analytic scale operationalizing four dimensions: (1) Task Achievement and Coherence, (2) Grammatical Range and Accuracy, (3) Lexical Resource, and (4) Mechanics (punctuation, spelling, paragraphing).

3.10. VOCABULARY ASSESSMENT PROTOCOL

Lexical competence was assessed using the Vocabulary Levels Test, a well-validated diagnostic instrument. Estimated active vocabulary size served as the primary dependent variable for lexical competence.

4. RESULTS AND DISCUSSION

4.1. SPEAKING SKILLS AND WRITING

The experimental group achieved a statistically significant improvement in speaking skills compared to the control group ($M_{EG} = 6.84$ vs. $M_{CG} = 5.97$ on the IELTS scale; $d = 0.71$; $p < 0.001$). Qualitative analysis revealed that participants in the experimental group developed greater fluency and a broader repertoire of lexical patterns, attributed to intensive conversational practice with ChatGPT without the pressure of public performance.

In the area of writing, grammatical accuracy increased by 24% in the experimental group, compared to 11% in the control group. ChatGPT's ability to provide immediate, analytically structured feedback on each written text proved particularly effective: participants received, on average, 3.2 additional corrective sessions per week compared to what would have been possible in traditional instruction.

These results are consistent with findings from prior research. A study conducted with 93 Chinese EFL students demonstrated that AI-supported instruction leads to superior learning outcomes and higher levels of engagement compared to traditional teaching.[5]

4.2. LEXICAL COMPETENCE

On the Vocabulary Levels Test, the experimental group achieved an average improvement of 312 words in active vocabulary (± 87), compared to 194 words (± 102) in the control group. Analysis of interactions with ChatGPT revealed that participants spontaneously developed a strategy of seeking contextual explanations for new words, consistent with the principles of contextualized vocabulary acquisition.

4.3. EXPERIMENT LIMITATIONS

Several limitations were identified. The experiment did not track long-term retention of learned material and future studies should incorporate delayed post-testing 3-6 months after the intervention. Additionally, the sampling was not geopolitically diversified, all participants came from the same cultural context, which limits the generalizability of the findings. There is also a risk of the Hawthorne effect. Finally, ChatGPT as a general-purpose platform occasionally generates inaccurate grammatical advice or stylistically inappropriate examples, requiring ongoing teacher supervision, a limitation not present in purpose-built ITS systems.

5. ETHICAL CHALLENGES AND IMPLEMENTATION LIMITATIONS

The use of ChatGPT in an educational context raises specific ethical challenges. Users may be reluctant to engage with AI language learning tools if they fear that their data will be misused or insufficiently protected. [6] In this experiment, participants were advised not to share personal information in conversations with ChatGPT, and sessions were conducted through educational accounts without the collection of personal data.

A particular challenge is the issue of academic integrity: ChatGPT can generate complete texts rather than helping participants develop their own competences. This risk was addressed through an explicit pedagogical protocol defining permissible and impermissible uses of the platform, along with regular discussions on the ethical use of AI in learning.

A third issue concerns the accuracy of information: ChatGPT may provide incorrect grammatical explanations or culturally inappropriate examples. The teacher was required to carry out weekly reviews of interaction logs and correct identified errors at the group level.

AI is not a replacement for the teacher, it rather transforms the teacher's role towards facilitating deeper understanding and cultural sensitivity.

6. FUTURE DIRECTIONS

Further development is anticipated in several key areas. First, advanced ChatGPT versions with multimodal capabilities (speech, image, and video analysis) open up the possibility of more comprehensive language practice encompassing both receptive and productive skills within a single environment. Second, the development of Custom GPT configurations enables teachers to create purpose-configured ChatGPT instances with specific pedagogical instructions, exercise sets, and evaluation criteria, thereby bridging the gap between a general-purpose platform and a specialized ITS system.

Third, emotionally adaptive learning, systems that recognize learner stress, frustration, or heightened motivation and dynamically adjust their approach represents a research direction of high potential. A systematic review indicates that the integration of generative AI is most advanced in the domain of writing and in higher education contexts, while speaking development, listening, and K-12 education are still in earlier stages of research maturity.[6]

7. CONCLUSION

The results of the experiment with 120 participants confirm statistically and pedagogically significant advantages of applying the ChatGPT platform compared to exclusively traditional instruction in English as a foreign language: speaking skills improved by 14.5%, writing accuracy by 13%, and lexical competence yielded 60% more new words in active vocabulary than in the control group. Motivation and self-regulated learning were statistically significantly higher among participants who used ChatGPT.

Unlike purpose-built ITS platforms, ChatGPT offers the advantage of broad accessibility and flexibility, along with inherent risks that require active teacher supervision and a clearly defined pedagogical protocol. These findings are consistent with current literature and meta-analytic evidence of a positive effect of AI on language skill development (g = 0.74).[3]

Future research should address long-term knowledge retention, the cultural diversity of samples, the development of ethical frameworks for participant data protection, and a comparison of the effectiveness of ChatGPT with dedicated ITS solutions. AI is not a replacement for pedagogical expertise and the emotional intelligence of the teacher but rather a powerful tool which, when integrated into a well-designed educational context, has the potential to democratize access to quality language education.

REFERENCES

[1] B. Klimova, M. Pikhart, and L. H. Al-Obaydi, "Exploring the potential of ChatGPT for foreign language education at the university level," *Front. Psychol.*, vol. 15, no. April, pp. 1–10, 2024, doi: 10.3389/fpsyg.2024.1269319.

[2] W. Wiboolyasarin, K. Wiboolyasarin, P. Tiranant, N. Jinowat, and P. Boonyakitanont, "AI-driven chatbots in second language education: A systematic review of their efficacy and pedagogical implications," *Ampersand*, vol. 14, no. May, p. 100224, 2025, doi: 10.1016/j.amper.2025.100224.

[3] Y. Doğan and T. Talan, "Artificial intelligence in foreign language learning: A bibliometric analysis," *J. Pedagog. Res.*, vol. 9, no. 2, pp. 206–230, 2025, doi: 10.33902/JPR.202427734.

[4] E. Alhusaiyan, "A systematic review of current trends in artificial intelligence in foreign language learning," *Saudi J. Lang. Stud.*, vol. 5, no. 1, pp. 1–16, 2025, doi: 10.1108/sjls-07-2024-0039.

[5] H. Qiao and A. Zhao, "Artificial intelligence-based language learning: illuminating the impact on speaking skills and self-regulation in Chinese EFL context," *Front. Psychol.*, vol. 14, no. November, 2023, doi: 10.3389/fpsyg.2023.1255594.

[6] B. Li, Y. L. Tan, C. Wang, and V. Lowell, "Two years of innovation: A systematic review of empirical generative AI research in language learning and teaching, "*Comput. Educ. Artif. Intell.*, vol. 9, no. July, p. 100445, 2025, doi: 10.1016/j.caeai.2025.100445.