Ed Process Int J

www.edupij.com

ISSN 2147-0901 (Print), 2564-8020 (Online)

EDUCATIONAL PROCESS
INTERNATIONAL JOURNAL

GenAI as a Learning Buddy for Non-English Majors: Effects on Listening and Writing Performance

Raigul Zheldibayeva

Research Article

Cite this article: Zheldibayeva, R. (2025). GenAI as a Learning Buddy for Non-English Majors: Effects on Listening and Writing Performance. *Educational Process: International Journal, 14*, e2025051. https://doi.org/10.22521/edupij.2025.14.51

Received January 01, 2025
Accepted February 04, 2025
Published Online February 12, 2025

Keywords:
Artificial intelligence, chatbots, ChatGPT, EFL students, Gemini.

Author for correspondence:
Raigul Zheldibayeva
r.zheldibayeva@zu.edu.kz
Department of Pedagogy and Psychology, Zhetysu University, 040000 Taldykorgan, Kazakhstan.

Abstract

Background/purpose. Despite the rapid assimilation of generative artificial intelligence (GenAI) into education, and initial evidence suggesting its potential in language learning, rigorous empirical investigations into its efficacy for cultivating specific language skills remain limited. This study aimed to address this lacuna by examining the impact of guided chatbot interactions on English as a Foreign Language (EFL) learners' listening and writing skills.

Materials/methods. A quasi-experimental design was employed, involving 93 undergraduates enrolled in mandatory EFL courses at a public university. Participants were randomly allocated to an experimental group (n=48), engaging in 10 weekly researcher-designed GenAI-mediated listening and writing activities, or a comparison group (n=45) following a traditional curriculum with optional chatbot use. Data were collected at three time points: pre-test, immediate post-test, and a delayed follow-up. Additionally, post-test qualitative reflections on chatbot use were garnered.

Results. While both groups demonstrated improvement over time, the experimental group reached significantly greater gains in both listening and writing at the immediate post-test compared to the control group. However, these advantages were not maintained at the follow-up assessment. Thematic analysis revealed that students valued the personalized and immediate feedback offered by the chatbots, yet expressed concerns about inconsistent content quality, occasional repetition, and the need for clearer task structures.

Conclusion. Overall, the findings suggest that targeted GenAI-mediated EFL activities can facilitate short-term improvements in listening and writing performance. Future research is advised to investigate approaches for sustaining these gains, optimizing content generation, and refining the user experience to better support second language learners' long-term development.

1. Introduction

Artificial intelligence (AI) has been a common thread weaving through nearly all technological advancements in the last circa half-century. Broadly, AI refers to systems and machines designed to replicate human cognitive processes, such as synthesizing, often to complete tasks typically associated with human intelligence (Kalniņa et al., 2024). From the pioneering "thinking machine" developed by Allen Newell and Herbert Simon in 1956, which was the first computer program designed to mimic human problem-solving abilities (Kohnke et al., 2023), to subsequent breakthroughs like the chess playing system Deep Blue and the virtual assistant Siri, AI systems have consistently demonstrated their capacity to imitate human expertise in specific domains. While being monumental for their time, these inventions primarily functioned as tools to exalt decision-making without generating entirely new content (García-López et al., 2024).

In recent years, however, the inception of generative AI (GenAI), exemplified by responsive chatbots such as ChatGPT, has heralded a paradigm shift in AI. This new class of AI stands out for its ability to instantly produce original, contextually relevant output by analyzing input data and identifying intricate patterns (Tafazoli, 2024). The modus operandi of these natural language systems is straightforward: users input their requests, commonly known as prompts, into a dialog box and rapidly receive a response, mostly in the form of text and/or images.

This transformative technology has swiftly permeated various sectors, including education. The education sector has readily embraced GenAI, incorporating it into learning management systems (Alier et al., 2025) and utilizing it for concept clarification, assessment, idea generation, etc. For many educators and students, GenAI represents their first meaningful encounter with AI technology (Parker et al., 2024). By providing immediate, personalized feedback and facilitating self-directed learning, GenAI tools hold the potential to alleviate the burden of teachers, while aiding students in cultivating essential skills and achieving a deeper understanding of their subjects (Rasul et al., 2024).

2. Literature Review

Summative evidence shows that although there has been a substantial accumulation of published papers and reviews on AI latterly, experiential non-opinion research specifically into the practical application of GenAI in language education is yet quite scarce (Law, 2024). A recent systematic review, encompassing 36 records, yielded a mere two empirical studies that reported students' language gains with quantifiable outcomes in the context of GenAI usage (Li et al., 2024). One of them is a single-case study (Li et al., 2023) involving Chinese language learners with varying levels of proficiency who engaged with ChatGPT for approximately 20 minutes twice a week at home. Then, there was a reversal period where the use of the chatbot was discontinued to monitor changes in their Chinese writing scores. The chatbot offered immediate feedback, error corrections, and helped to compose well-structured sentences, guiding the learners in their writing assignments. Eventually, all students experienced significant enhancements in their Chinese writing scores during both the GenAI assistance and subsequent withdrawal phases. Another study (Escalante et al., 2023) revealed no significant difference in second language (L2) attainment between university EFL students who leveraged support from ChatGPT and those who received feedback from human tutors.

The literature search uncovered one additional investigation (Shahsavar et al., 2024) in which integrating a chatbot into medical students' writing activities resulted in more noticeable advancements in their English academic writing skills relative to the counterparts who followed traditional writing instruction.

As for the development of L2 speaking skills through GenAI, only one relevant GenAI intervention that could be found to date is a quasi-experiment (Chen et al., 2024) in which non-English speaking students learning English used a GenAI agent for role-playing. This method was effective in elevating

https://doi.org/10.22521/edupij.2025.14.51 Published online by Universitepark Press

their oral performance, but the improvements were no greater than those seen in a reference group utilizing conventional peer-to-peer role-playing. At the time of writing, no prior experimental studies with clear quantitative outcomes on L2 listening skills could be retrieved from the academic literature.

2.1. Problem

This concise literature review corroborates a previously expressed concern that the impact of GenAI has been underexplored through experimentations (Yusuf et al., 2024). Furthermore, the existing research body is evidently skewed towards the topic of writing skills. Despite the recent surge in the popularity of the generative technology, its deployment is still in its early phases, and the research landscape remains similarly immature, necessitating further investigation into how chatbots can promote the educational process (Foung et al., 2024). The prevailing sentiment surrounding GenAI indicates a growing recognition of the need for reforms to better assimilate the technology, yet there is an urgent demand for research to guide educators, students, and curriculum designers in effectively harnessing GenAI in education (Xia et al., 2024; Zhai, 2024). Hence, further investigation is required to test the potential to facilitate foreign language learning using the multimodal capabilities of today's conversational agents.

Despite the notable deficit of research focused on enhancing speaking and listening abilities using generative technologies, the present work does not address speaking skills, as L2 oral performance is heavily contingent upon factors such as personality traits (Kim & Hwang, 2024), speaking anxiety (Mora et al., 2024), and interlocutor characteristics (Crowther & Isbell, 2023). These variables are beyond the scope of what a comprehensive technology-centered intervention can effectively encompass, and their presence would hinder researchers' ability to isolate the specific effects of GenAI. The scant research on generative chatbots as a means of EFL speaking attainment has thus far failed to evince significant impacts on speaking skills compared to interactions with teachers and peers (Chen et al., 2024; Wang et al., 2024). This study instead concentrates on domains, namely writing and listening skills, where students are more likely to benefit from the textual and auditory functionalities of the smart content-generating assistants. Chatbots can generate level-appropriate listening materials, which can then be read aloud using built-in text-to-speech functions available in some models. Learners thereby can engage with both auditory output and written text, facilitating the cultivation of phonological awareness, vocabulary recognition, and listening comprehension. Moreover, the ability to replay the audio and interact with comprehension tasks proposed by the virtual agent can help students reinforce their understanding and monitor their listening progress.

Similarly, for writing skills, GenAI systems provide immediate, targeted feedback on grammar, structure, and coherence - elements essential for developing well-formed written communication. By meaningfully engaging with generative tools, students can iteratively revise their work, glean progress points, and internalize key principles of academic writing (Jackaria et al., 2024; Marzuki et al., 2023). This type of real-time, low-pressure feedback can be particularly beneficial for learners who may lack confidence or access to human instructors.

2.2. Study Goal and Relevance

To address the highlighted evidence gap, this research seeks to examine the impacts of chatbot-assisted EFL learning on students' English proficiency. Specifically, three research questions (RQs) were propounded:

RQ1. How does weekly interaction with chatbots, compared to traditional learning activities, influence students' listening performance?

https://doi.org/10.22521/edupij.2025.14.51 Published online by Universitepark Press

RQ2. How does weekly interaction with chatbots, compared to traditional learning activities, influence students' writing performance?

By tackling these questions, the study endeavors to explore how contemporary generative agents can aid students in acquiring two critical language skills: the ability to comprehend spoken language and the ability to articulate their thoughts and opinions in written form.

RQ3. What are the strengths, challenges, and recommendations expressed by students regarding their experiences with using chatbots for English learning?

Through this question, it is expected to gather revelations into the perceived benefits and pitfalls of GenAI from the participants' perspective, along with suggestions for its successful incorporation into language education.

Revelations from this inquiry will hopefully guide practitioners in embedding GenAI effectively into language education frameworks, eventually enriching the learning experience and outcomes for students. The subsequent Methods section details the manipulations designed to answer these research questions.

3. Methodology

3.1. Research Design

This study employed a convergent parallel mixed-methods design. The quantitative component utilized a quasi-experimental pretest-posttest-follow-up design with a control group. The qualitative component involved gathering participants' perceptions. The integration of quantitative and qualitative data allowed for a comprehensive understanding of the impact of the intervention on students' listening and writing skills, as well as their perspectives on the use of GenAI.

3.2. Participants

The target population comprised undergraduate students at a public university in [blinded for review], pursuing non-English-major degrees and enrolled in mandatory EFL courses. Ethics committee approval was obtained from the first author's institution. A purposeful sampling approach was utilized, targeting all students in designated EFL classes. Initially, 121 students consented to participate after receiving detailed information about the study's procedures, their rights (including the right to withdraw without penalty), and data confidentiality. After excluding 28 individuals who failed to complete all required procedures, the final sample included 93 participants (58 females and 35 males, aged 18-23) enrolled in Bachelor programs across the Arts and Humanities, Services, Education, or Agriculture departments. All participants provided informed consent prior to commencement. The sample's English proficiency, based on university records, was estimated to be around the B1 level of the Common European Framework of Reference (CEFR).

3.3. Intervention

Before the baseline evaluation, the 93 participants were randomly assigned to either a comparison group (n = 45) supposed to follow a traditional curriculum with optional chatbot use or an experimental group (n = 48) supposed to engage in ten consecutive weekly GenAI-mediated learning activities. The ten-week intervention commenced in early March 2024. The experimental condition involved: (1) Gemini for Listening. Each week, students applied a researcher-designed listening comprehension prompt (Appendix I) to trigger Google's Gemini to generate a level-appropriate story or conversation. By pressing a "Listen" button, students could hear the text without reading the on-screen text. The chatbot then presented a comprehension task, checked responses and offered further text with corresponding exercise; (2) ChatGPT for Writing. Students also interacted with a private ChatGPT-powered Telegram chatbot (configured via BotFather using an

https://doi.org/10.22521/edupij.2025.14.51 Published online by Universitepark Press

OpenAI API key), designed by the author to provide feedback on academic writing assignments based on researcher-suggested topics. This system employed a hidden prompt that instructed the AI to serve as an academic writing instructor. Each week, participants selected a topic from a researcher-provided list, wrote at least 300 words, and submitted this text to the chatbot for feedback on coherence, structure, grammar, and clarity. Based on these criteria, GenAI scored the assignment from 0 ("The writing is blank, rejects the topic, or is not in English") to 5 ("The writing is a relevant and very clearly expressed contribution to the discussion, with well-elaborated explanations, effective use of syntactic structures, and minimal errors") in accordance with TOEFL iBT (Test of English as a Foreign Language Internet-based Test) and generated recommendations. Students were encouraged to complete these activities at their convenience, but weekly engagement was mandatory, confirmed through screenshots submitted to an anonymous research assistant.

Comparison group students followed the same weekly instructional schedule in the classroom (e.g., worksheet-based activities, audiotape listening exercises, and writing assignments). To balance the treatment group's increased learning opportunities, the reference group was offered to utilize chatbots of their choice for bolstering their language skills, albeit without access to the specific prompts and tailored listening practice chatbot provided to the experimental group. This arrangement aimed to control for the potential effect of simply using a chatbot, isolating the impact of the researcher-designed manipulations.

Both groups possessed self-reported moderate prior experience with GenAI and were provided the link to an online course on how to use GenAI for academic purposes (teachgenai.au.dk/learn-genai/learn-genai-course) covering functionalities (including prompting skills), limitations, and ethical considerations. This ensured a baseline level of GenAI familiarity across both groups.

3.4. Instruments and Assessments

Three assessment time points were employed: (a) pre-test in late February 2024, (b) immediate post-test in mid-May 2024, and (c) delayed follow-up in early September 2024, after a summer break.

Listening comprehension was assessed using paper-pencil cloze tests. Research assistants, blind to group allocation, played audio recordings of conversational or story vignettes, followed by a cloze task with ten contextually unpredictable omitted words (no options provided). Scores reflected the number of correctly identified words. Task difficulty was rigorously maintained across the three administrations.

Writing performance was evaluated through 45-minute assignments on topics derived from an earlier study (Han, 2024) for the pre- and immediate post-tests to ensure prior knowledge was not a factor. A similar self-constructed topic was employed for the follow-up (In today's interconnected world, do you believe that globalization has primarily fostered cooperation and understanding, or has it led to increased conflict and division? Support your answer with specific examples and details). The topics focused on debatable issues requiring reasoned arguments and specific examples. All writing samples were scored using the academic discussion rubric (ets.org/pdfs/toefl/toefl-ibt-writing-rubrics.pdf) from the writing section of the TOEFL iBT (Test of English as a Foreign Language Internet-based Test), ensuring a reliable and valid measure of performance. Two English teachers (PhD holders, each with over 10 years of teaching, unaware of group assignment, affiliated to universities beyond the study site) independently scored all listening and writing assessments. Inter-rater reliability was ensured via simple percentage agreement (above 80%). Where there was a mismatch, the average score of both raters was calculated.

3.5. Qualitative Data Collection and Analysis

The immediate post-test for the experimental group included a free-response form with three questions adapted from Karataş et al. (2024), inviting participants to reflect (in their native language)

https://doi.org/10.22521/edupij.2025.14.51 Published online by Universitepark Press

on the beneficial aspects, limitations, and potential improvements of using the conversational agents in language learning. Responses were analyzed using a primarily deductive coding scheme with emergent thematic analysis. Three a priori categories - strengths, challenges, and recommendations - guided initial coding, while themes emerged inductively within each category. Two trained coders independently reviewed the responses, identifying emergent themes through discussion and consensus. Translated excerpts of the responses were included in the manuscript for reporting.

3.6. Quantitative Data Analysis

To examine the effects of the intervention over time, quantitative data (listening and writing scores) were analyzed via repeated-measures analysis of variance (ANOVA) by applying R software packages. Holm-adjusted paired and unpaired t-tests probed intragroup and intergroup differences, respectively. Statistical significance was set at $\alpha < 0.05$. Assumptions of normality, sphericity, and homogeneity of variances were verified using Q-Q plots, Mauchly's test, and Levene's test, respectively, and were not violated.

3.7. Pilot Study

Prior to full implementation, a pilot study with eight non-English-majoring students was conducted to determine the feasibility of the intervention algorithm along with the assessment, coding and scoring procedures. Minor inconveniences and technical hitches were encountered, particularly in verifying engagement by emailing screenshots and ensuring participants' privacy. As a result, the researcher developed more explicit instructions for participants to remove all personally identifying details from their screenshots before emailing them. No other critical modifications to the intervention or assessment procedures were required.

4. Results

4.1. Quantitative Analysis

Listening Scores

The results of the repeated-measures ANOVA revealed a significant main effect of time ($F(2, 182) = 93.14$, $p = 0.001$, $\eta 2 = 0.168$), group ($F(1, 91) = 5.58$, $p = 0.02$, $\eta 2 = 0.047$), and group-by-time interaction ($F(2, 182) = 9.39$, $p = 0.001$, $\eta 2 = 0.020$). These findings indicate that listening scores increased over time for both groups, but the experimental group demonstrated a significantly greater progression compared to the control group.

At pre-test, there were no significant differences between the control group ($M = 4.20$, $SD = 1.10$) and the experimental group ($M = 4.44$, $SD = 1.25$) ($t(91) = -0.97$, $p = .415$). Nonetheless, at the post-test, the experimental group ($M = 6.21$, $SD = 1.68$) significantly surpassed the control group ($M = 5.09$, $SD = 1.35$) ($t(89) = -3.56$, $p = 0.002$). This advantage was somewhat reduced at the follow-up assessment, where scores for the experimental group declined to $M = 5.73$ ($SD = 1.48$), while the control group advanced slightly to $M = 5.31$ ($SD = 1.10$). The difference at follow-up was not statistically significant ($t(87) = -1.55$, $p = 0.376$).

Within-group comparisons yielded substantial gains for both groups throughout the intervention. In the experimental group, listening scores rose significantly from time 1 to time 2 ($t(47) = -12.61$, $p = 0.001$) and remained significantly higher at follow-up ($t(47) = -9.26$, $p = 0.001$), though performance dropped slightly from post-test to follow-up ($t(47) = 4.02$, $p = 0.001$). The control group also exhibited significant enhancements from pre-test to post-test ($t(44) = -5.26$, $p = 0.001$) and pre-test to follow-up ($t(44) = -6.58$, $p = 0.001$), but no significant change was observed between post-test

https://doi.org/10.22521/edupij.2025.14.51 Published online by Universitepark Press

and follow-up (t(44) = -1.28, p = 0.415). Figure 1 provides a visual summary of the listening scores for both groups across the three evaluations.

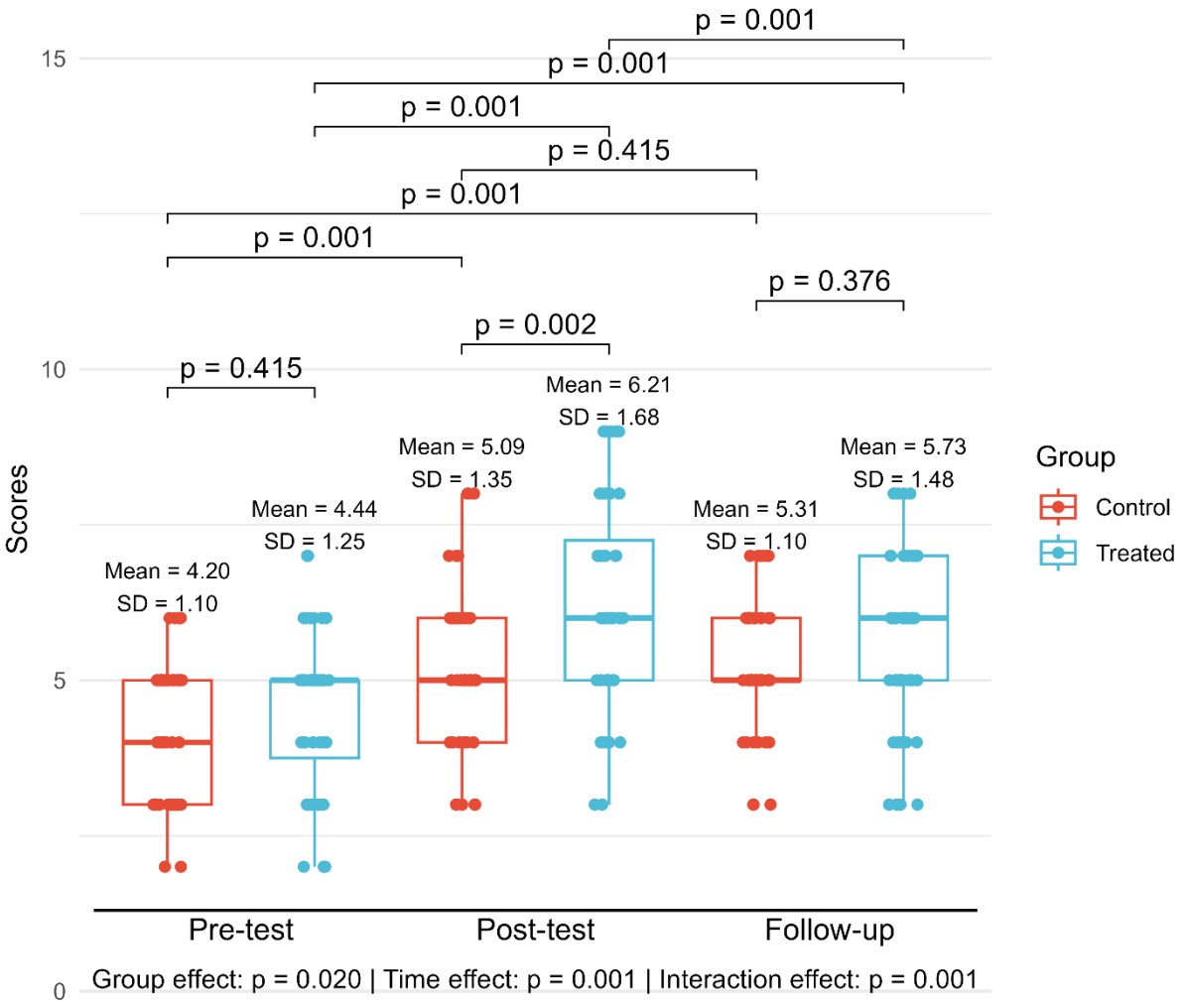

Figure 1. Listening scores. Boxplots represent median and interquartile range. Below boxplots are RM ANOVA p-values, above boxplots are Holm-corrected t-test p-values.

Writing Scores

For writing scores, the repeated-measures ANOVA revealed a significant main effect of time (F(2, 182) = 20.51, p = 0.001, η2 = 0.068) and group*time interaction (F(2, 182) = 8.62, p = 0.001, η2 = 0.030). However, the main effect of group was insignificant (F(1, 91) = 1.62, p = 0.206), suggesting that overall, differences between the experimental and control groups were modest, though improvements over time were more pronounced in the experimental group.

At pre-test, the control group (M = 3.02, SD = 0.66) and experimental group (M = 2.90, SD = 0.69) showed no significant difference (t(91) = 0.90, p = 1.00). By post-test, the experimental group attained a significantly higher mean score (M = 3.63, SD = 0.73) in contrast to the control group (M = 3.18, SD = 0.65) (t(91) = -3.12, p = 0.017). At follow-up, writing scores slightly decreased for both groups. The experimental group scored M = 3.21 (SD = 0.62), while the control group scored M = 3.09 (SD = 0.70), with no significant difference between the two groups (t(88) = -0.87, p = 1.00).

Inter-group comparisons detected that the experimental group had a significant advancement from pre-assessment to post-assessment (t(47) = -7.85, p = 0.001), and the gains were marginally sustained at follow-up relative to baseline (t(47) = -2.70, p = 0.058). However, a significant drop was observed from post-test to follow-up (t(47) = 4.71, p = 0.001). Conversely, the control group demonstrated no significant progress at any time point, with pre-evaluation, post-evaluation, and

https://doi.org/10.22521/edupij.2025.14.51 Published online by Universitepark Press

follow-up scores remaining statistically similar. Figure 2 illustrates the trends in writing scores across the three time points for both groups.

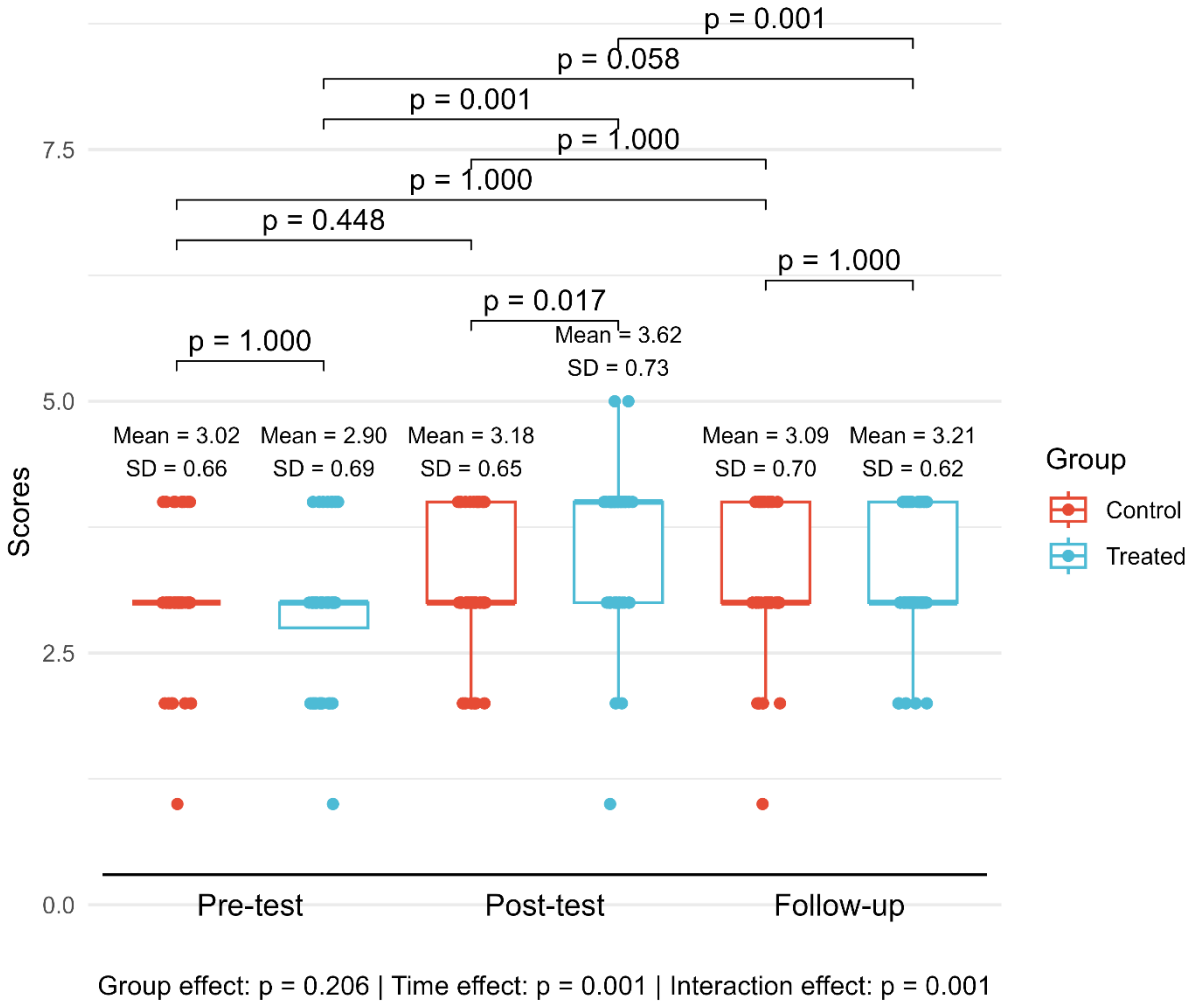

Figure 2. Writing scores. Boxplots represent median and interquartile range. Below boxplots are RM ANOVA p-values, above boxplots are Holm-corrected t-test p-values.

4.2. Qualitative Analysis

Thematic analysis of the participants' experiences with using GenAI chatbots as learning companions yielded three overarching topics: strengths, challenges, and recommendations, within which several key themes emerged.

Strengths

Participants identified several positive aspects of using chatbots to support their English language learning. These benefits clustered around two main themes: personalized and engaging language exposure, and immediate and constructive feedback.

Theme 1: Personalized and Engaging Language Exposure

A significant advantage highlighted by participants was the chatbots' capacity to deliver personalized and engaging language input. This was particularly evident in the listening practice with Gemini, where the ability to generate stories and conversations based on learner interests or specific topics fostered greater motivation and focus. As Participant 14 articulated, "Gemini let me choose topics I actually cared about, like gaming or travel, so I was more motivated to listen and focus. It was not just random stuff I had no interest in, like some of the textbook examples." This personalization extended to academic needs, with Participant 27 noting, "When I asked Gemini to generate stories

https://doi.org/10.22521/edupij.2025.14.51 Published online by Universitepark Press

about environmental science, which is my major, it created texts that helped me learn field-specific vocabulary while practicing listening." The dynamic nature of the content generation also encouraged active listening, as noted by Participant 8: "Having the bot generate new stories for each session meant I was not just memorizing answers. I had to really listen and understand."

The writing chatbot also facilitated a more personalized learning experience. Participants appreciated the targeted feedback that identified recurring patterns in their writing. For instance, Participant 33 shared, "What I liked in the writing bot is it spotted patterns in my essays. Like when I kept using 'however' at the beginning of sentences, it suggested different ways to vary my transitions." The feedback also fostered greater awareness of academic writing conventions, with Participant 5 stating, "I really liked how the Telegram bot gave me specific advice on my writing. Rather than being general comments, it pointed out exactly where I could improve, like using more transition words between paragraphs." This focus on specific areas for improvement helped students become more aware of their writing choices. One participant explained, "I did not realize before that I should directly address the question in every paragraph. The chatbot reminded me to stay focused on the topic and not add unnecessary details" (Participant 20). Another highlighted the value of vocabulary refinement, stating, "It pointed out that I was repeating words like 'important' too much. I tried replacing them with more academic words like 'significant' or 'crucial,' and it made my writing better" (Participant 12).

Theme 2: Immediate and Constructive Feedback

While the novelty of instant feedback might be common for digital natives, participants emphasized its practicality and, importantly, its constructive nature. This was particularly valued in the writing tasks, where the Telegram chatbot's feedback was perceived as actionable and aligned with academic writing standards. As Participant 25 explained, "The feedback was not just 'this is wrong' – it told me exactly how to fix it. Like, it said I should use more linking words to connect my ideas, and even gave me examples. That was way more useful than just getting a grade." Similarly, in the listening exercises, the immediate evaluation of cloze tasks by Gemini provided a direct measure of comprehension, allowing for dynamic adjustments to the learning process. Participant 41 noted, "After I answered the cloze questions, Gemini told me right away if I got them right or wrong. If I messed up, it gave me another story on the same topic, which made me feel like I was actually improving."

Challenges

Despite the perceived benefits, participants also identified challenges associated with the use of chatbots for language learning. These challenges primarily centered on the inconsistent quality of generated content and the potential for repetition in listening tasks.

Theme 1: Inconsistent Quality of Generated Content

A recurring concern was the variability in the quality of content generated by Gemini. Some participants found the stories and conversations to be occasionally simplistic or lacking in depth, which impacted their engagement. Participant 19 commented, "Sometimes the stories from Gemini felt a bit childish, like they were written for kids. It was hard to stay focused when the content did not challenge me enough." Issues with the cloze task design were also noted, as highlighted by Participant 36: "Sometimes, Gemini would remove words from the story that were actually quite easy to guess from the context."

Theme 2: Repetition in Listening Tasks

While the adaptive mechanism of providing another story upon failing the initial cloze task was intended to support learning, some participants found it repetitive. Participant 19 explained, "If I did not get half the answers right, the chatbot made me do another story on the same topic. It was

https://doi.org/10.22521/edupij.2025.14.51 Published online by Universitepark Press

helpful, but after a while, it felt repetitive, especially when I was tired or did not feel like focusing on the same thing again." This suggests that while the adaptive feature is beneficial, strategies to mitigate potential monotony might be necessary.

Recommendations

Based on their experiences, participants offered several recommendations to enhance the effectiveness and user-friendliness of chatbots for language learning. These suggestions focused on simplifying and structuring writing feedback, improving listening task usability, and enhancing listening task consistency and variety.

Theme 1: Simplifying and Structuring Writing Feedback

Participants suggested that the writing feedback provided by the chatbot could be more easily digestible by focusing on fewer key areas for improvement at a time. Participant 15 proposed, "Perhaps the feedback could be simpler and more focused. For example, instead of listing multiple areas to improve, it could focus on one or two key points per assignment. This would make it easier to apply."

Theme 2: Improving Listening Task Usability

Several recommendations aimed at improving the practicality of the listening tasks. Suggestions included streamlining the interface and enhancing audio playback controls. Participant 29 suggested, "It would be great if the chatbot automatically played the audio instead of requiring me to press the 'Listen' button." Another participant highlighted the need for clearer visual cues during audio playback, stating, "Hopefully, later iterations of the chatbot will feature a visual cue, like a progress bar, to show when the audio is playing, so I know if it is working properly" (Participant 43).

Theme 3: Enhancing Listening Task Consistency and Variety

To further optimize the listening practice, participants recommended measures to ensure consistent difficulty levels and a greater diversity of content. Participant 29 suggested, "I would prefer that the chatbot could ensure the stories are always at the actual difficulty level. Maybe it could ask me to rate the difficulty after each story, so it can adjust for the next session."

To resume, participants appreciated the personalized learning experiences and immediate feedback. However, they also pointed out areas for improvement, particularly regarding content consistency and task usability.

5. Discussion

Essential findings worth discussing from the research include the fact that the learning curriculum in schools may be very complicated and not easy to implement. Teachers developing a new learning model must have the courage to simplify learning objectives. The teacher's ability to express learning objectives and theme-taking is highly recommended (Kulhmann Lüdeke & Sánchez Zúñiga, 2017; Nurmadiah et al., 2022) so that teachers are more independent when choosing a new learning model. The ability of teachers to formulate teaching objectives and materials is vital. This requires teachers to be more professional in teaching (Cai Zhaohui, 2014; Nor et al., 2022).

This investigation set out to scrutinize whether the integration of chatbots (Gemini for listening practice and ChatGPT for writing tasks) could bolster English as a Foreign Language (EFL) students' proficiency in listening and writing. Specifically, it addressed three questions: how chatbot-mediated activities compared to traditional activities would influence students' listening (RQ1) and writing (RQ2), and how learners perceived the strengths, challenges, and future possibilities of these technological tools (RQ3). In summary, the quantitative findings showed that learners who engaged weekly with these generative AI tools improved their listening and writing performance more

https://doi.org/10.22521/edupij.2025.14.51 Published online by Universitepark Press

significantly from baseline to immediate post-test than those in the control group. Nevertheless, both listening and writing scores partially regressed for the experimental group by the delayed follow-up. Qualitative data complemented these results by highlighting the personalized, immediate feedback afforded by chatbots, while also pointing to concerns related to inconsistent AI-generated content quality and repetitive listening tasks. Taken together, these findings affirm the potential educational value of generative AI agents in facilitating language learning, although sustaining those gains over time remains a challenge.

The results regarding ChatGPT's contribution to writing enhancement bear partial congruence with the positive impact of GenAI on writing, documented in studies like Shahsavar et al. (2024). The advancement in writing proficiency observed in the treatment group aligns with the general trend of improvement reported in some previous research, though nuanced distinctions arise. Song and Song (2023) similarly reported marked improvements in writing competence, noting gains in organization, coherence, and grammar, which resonate with the present study's results that the experimental group significantly outperformed the control group at the immediate post-test. In contrast, Escalante et al. (2023) found no significant difference between AI-generated writing feedback and human tutors; the current study does diverge by showing ChatGPT-based instruction can yield short-term advantages in writing performance (though not sustained longitudinally). These contradictory results may be explained by differences in treatment duration, feedback protocols, or learner profiles across the studies. Further, Ironsi and Ironsi (2024) emphasized that ChatGPT is beneficial in terms of generating ideas but might not thoroughly fortify overall writing skills. A partial echo of this perspective can be observed in the current findings, wherein initial improvements at post-test were not robustly upheld at follow-up. Meanwhile, qualitative investigations by Karataş et al. (2024) and Kim et al. (2024) highlight the perceived benefits of ChatGPT in providing swift, scaffolded feedback: the present qualitative results similarly show strong learner appreciation for the chatbot's dynamic pointers but also note apprehensions relating to potential over-reliance. These parallels suggest that while learners may experience immediate motivational and performance gains, lasting effects demand further scaffolding and varied practice. The observed improvements in writing scores within the experimental group can be attributed to the fact that the bespoke chatbot provided students with immediate and targeted feedback on various aspects of their writing in line with a globally recognized assessment system. The bot's capacity to recognize patterns in students' writing allowed for focused attention on recurring issues, potentially leading to more profound improvements in coherence, structure, and grammatical accuracy.

As for the listening domain, this study ventures into a territory largely uncharted by past research. The complete absence of experimental studies with measurable outcomes on L2 listening skills, as highlighted in the literature review, underscores the novelty of this investigation. The observed gains in the experimental group's listening scores, exceeding those of the control group, signify the potential of tailored, interactive listening exercises generated by AI. Gemini's adaptivity and ability to generate learner-specific narratives likely fostered more engaged, contextually meaningful listening experiences. By tailoring content to learners' fields of interest, Gemini may have stimulated heightened motivation and attentional focus, both of which are instrumental in listening acquisition. Moreover, the iterative nature of the listening tasks, with immediate feedback and opportunities for repeated exposure to similar content, may have solidified comprehension and reinforced vocabulary acquisition. However, there is a dearth of research employing this particular generative agent for listening comprehension development, rendering direct comparisons with past interventions impracticable. Some scholars (e.g., Imran & Almusharraf, 2024) have underscored Gemini's promise for educational contexts, yet empirical endeavors frequently overlook it in favor of ChatGPT. It is hoped that the present study can serve as a catalyst, encouraging researchers to explore the untapped potential of Gemini in diverse educational settings. The qualitative data further

https://doi.org/10.22521/edupij.2025.14.51 Published online by Universitepark Press

supports these quantitative findings, with students explicitly mentioning the benefits of personalized content for listening and the actionable nature of the writing feedback.

The paucity of empirical studies noted in the literature review, particularly those with quantifiable outcomes, positions this research as a valuable contribution to a nascent field. The concentration of earlier work on writing interventions is also addressed by this study's dual focus on both writing and listening skills development.

5.1. Limitations and Future Research

Despite its contributions, this inquiry has certain constraints to acknowledge. One limitation concerns the relatively short intervention period of ten weeks, followed by a summer break that may have confounded retention effects. Next, the requirement for participants to submit screenshots as proof of chatbot interaction possibly disrupted the flow of their language practice. Future interventions could examine alternative verification methods for chatbot engagement that do not interrupt users' natural learning flow, potentially through unobtrusive tracking tools or integrated log data, e.g., like in Guo et al. (2024). Furthermore, the study did not isolate the effects of personalized interests in listening content from the chatbot-based modality itself, leaving open the question of whether similar results would emerge with non-personalized AI content. Consequently, similar investigations to come should try to disentangle the role of tailored content from the technology platform itself, for instance by comparing personalized listening materials with generic AI-driven tasks and measuring any resulting differences in motivation or performance. Lastly, participants' self-reported use of the chatbot outside the required tasks remains partially unmonitored, limiting the control over extraneous variability. Future studies may embed systematic monitoring strategies to account for participants' supplementary chatbot use, capturing more accurate data on the frequency, duration, and type of extracurricular interactions. By addressing these considerations, subsequent research can yield deeper insights into how personalized AI-based language practice functions in authentic contexts and provide more robust evidence for its efficacy.

The preliminary nature of this study inherently constrained its scope, precluding the ability to address all of the aforementioned considerations within a single investigation. As an initial exploration, its primary aim was to establish a foundational understanding that subsequent research could refine and build upon. By focusing on the central aspects of chatbot-based language practice, this study intentionally foregrounded the fundamental interactions and outcomes, thereby generating baseline data that can guide future inquiries. Consequently, the study did not undertake more detailed assessments - such as the isolated impact of personalized content or comprehensive monitoring mechanisms - because its primary purpose was to set forth a preliminary framework rather than provide conclusive evidence. This foundational step allows subsequent research to adopt more robust designs and employ additional controls, ensuring that emerging questions and methodological gaps can be addressed with greater precision and depth.

5.2. Recommendations

Building on these findings, several course-level and learner-level suggestions emerge for maximizing the benefits of AI tools. First, it is advisable for EFL learners to adopt purposeful strategies when employing chatbots, such as tailoring prompts to address specific linguistic needs and seeking variety in chatbot-generated tasks to avoid repetitive practice. Second, educators might want to incorporate structured reflection activities, prompting students to note which language features they struggle with and whether chatbot feedback directly helps them overcome these hurdles. Finally, the variety and difficulty level of AI-generated listening tasks should be regularly monitored and tweaked to sustain motivation and linguistic growth.

https://doi.org/10.22521/edupij.2025.14.51 Published online by Universitepark Press

6. Conclusion

To sum up, the investigation described here revealed evidence from practice for the pedagogical promise of GenAI in supporting EFL students' listening and writing skill development. This research is the first to assess the impact of Gemini-mediated learning on listening comprehension in an EFL context, thereby addressing a noticeable gap in the extant literature. By garnering both quantitative performance data and qualitative student perspectives, this paper sheds light on both benefits and pitfalls of chatbot use in language education. The findings can be assistive for educators seeking to thoughtfully implement generative solutions for L2 learning and teaching. This study was designed to be rigorous yet applicable, and it is the author's hope that it has contributed to a deeper comprehension of the intricate interaction between artificial intelligence, pedagogy, and the acquisition of foreign languages. Finally, it is paramount to continue to critically appraise the potential benefits and challenges of these emerging technologies as they increasingly become entwined with the fabric of education.

Declarations

Funding. This research was funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan under the project "Artificial intelligence as an assistant in acquiring academic English by non-language major students: Randomized study" (AP25794537).

References

Alier, M., Pereira, J., Garcia-Peñalvo, F. J., Casañ, M. J., & Cabré, J. (2025). LAMB: An open-source software framework to create artificial intelligence assistants deployed and integrated into learning management systems. *Computer Standards and Interfaces, 92*, 103940. https://doi.org/10.1016/j.csi.2024.103940

Chen, A., Jia, J., Li, Y., & Fu, L. (2024). Investigating the effect of role-play activity with GenAI agent on EFL students' speaking performance. *Journal of Educational Computing Research*. Advance online publication. https://doi.org/10.1177/07356331241299058

Crowther, D., & Isbell, D. R. (2023). Second language speech comprehensibility: A research agenda. *Language Teaching*. Advance online publication. https://doi.org/10.1017/s026144482300037x

Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education, 20*, 57. https://doi.org/10.1186/s41239-023-00425-2

Foung, D., Lin, L., & Chen, J. (2024). Reinventing assessments with ChatGPT and other online tools: Opportunities for GenAI-empowered assessment practices. *Computers and Education Artificial Intelligence, 6*, 100250. https://doi.org/10.1016/j.caeai.2024.100250

García-López, I. M., González, C. S. G., Ramírez-Montoya, M., & Molina-Espinosa, J. (2024). Challenges of implementing ChatGPT on education: Systematic literature review. *International Journal of Educational Research Open, 8*, 100401. https://doi.org/10.1016/j.ijedro.2024.100401

Guo, K., Di Zhang, E., Li, D., & Yu, S. (2024). Using AI-supported peer review to enhance feedback literacy: An investigation of students' revision of feedback on peers' essays. *British Journal of Educational Technology*. Advance online publication. https://doi.org/10.1111/bjet.13540

Han, L. (2024). Metacognitive writing strategy instruction in the EFL context: Focus on writing performance and motivation. *Sage Open, 14*(2). https://doi.org/10.1177/21582440241257081

Imran, M., & Almusharraf, N. (2024). Google Gemini as a next generation AI educational tool: A review of emerging educational technology. *Smart Learning Environments, 11*, 22. https://doi.org/10.1186/s40561-024-00310-z

Ironsi, C. S., & Ironsi, S. S. (2024). Experimental evidence for the efficacy of generative AI in improving students' writing skills. *Quality Assurance in Education*. Advance online publication. https://doi.org/10.1108/qae-04-2024-0065

Jackaria, P. M., Hajan, B. H., & Mastul, A. H. (2024). A comparative analysis of the rating of college students' essays by ChatGPT versus human raters. *International Journal of Learning Teaching and Educational Research, 23*(2), 478–492. https://doi.org/10.26803/ijlter.23.2.23

Kalniņa, D., Nīmante, D., & Baranova, S. (2024). Artificial intelligence for higher education: Benefits and challenges for pre-service teachers. *Frontiers in Education, 9*, 1501819. https://doi.org/10.3389/feduc.2024.1501819

Karataş, F., Abedi, F. Y., Gunyel, F. O., Karadeniz, D., & Kuzgun, Y. (2024). Incorporating AI in foreign language education: An investigation into ChatGPT's effect on foreign language learners. *Education and Information Technologies, 29*(15), 19343–19366. https://doi.org/10.1007/s10639-024-12574-6

Kim, J., & Hwang, Y. (2024). How are personality factors connected to EFL learners' oral performance? A psychological network analysis. *System, 127*, 103516. https://doi.org/10.1016/j.system.2024.103516

Kim, J., Yu, S., Detrick, R., & Li, N. (2024). Exploring students' perspectives on Generative AI-assisted academic writing. *Education and Information Technologies*. Advance online publication. https://doi.org/10.1007/s10639-024-12878-7

Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). Exploring generative artificial intelligence preparedness among university language instructors: A case study. *Computers and Education: Artificial Intelligence, 5*, 100156. https://doi.org/10.1016/j.caeai.2023.100156

Law, L. (2024). Application of generative artificial intelligence (GenAI) in language teaching and learning: A scoping literature review. *Computers and Education Open, 6*, 100174. https://doi.org/10.1016/j.caeo.2024.100174

Li, B., Lowell, V. L., Wang, C., & Li, X. (2024). A systematic review of the first year of publications on ChatGPT and language education: Examining research on ChatGPT's use in language learning and teaching. *Computers and Education Artificial Intelligence, 7*, 100266. https://doi.org/10.1016/j.caeai.2024.100266

Li, X., Li, B., & Cho, S. (2023). Empowering Chinese language learners from low-income families to improve their Chinese writing with ChatGPT's assistance afterschool. *Languages, 8*(4), 238. https://doi.org/10.3390/languages8040238

Lim, W. M., Gunasekara, A., Pallant, J. L., Pallant, J. I., & Pechenkina, E. (2023). Generative AI and the future of education: Ragnarök or reformation? A paradoxical perspective from management educators. *International Journal of Management Education, 21*(2), 100790. https://doi.org/10.1016/j.ijme.2023.100790

Marzuki, Widiati, U., Rusdin, D., Darwin, & Indrawati, I. (2023). The impact of AI writing tools on the content and organization of students' writing: EFL teachers' perspective. *Cogent Education, 10*(2), 2236469. https://doi.org/10.1080/2331186x.2023.2236469

Mora, J. C., Mora-Plaza, I., & Miranda, G. B. (2024). Speaking anxiety and task complexity effects on second language speech. *International Journal of Applied Linguistics, 34*(1), 292–315. https://doi.org/10.1111/ijal.12494

Parker, L., Carter, C., Karakas, A., Loper, A. J., & Sokkar, A. (2024). Graduate instructors navigating the AI frontier: The role of ChatGPT in higher education. *Computers and Education Open, 6*, 100166. https://doi.org/10.1016/j.caeo.2024.100166

Rasul, T., Nair, S., Kalendra, D., Balaji, De Oliveira Santini, F., Ladeira, W. J., Rather, R. A., Yasin, N., Rodriguez, R. V., Kokkalis, P., Murad, M. W., & Hossain, M. U. (2024). Enhancing academic integrity among students in GenAI Era: A holistic framework. *International Journal of Management Education, 22*(3), 101041. https://doi.org/10.1016/j.ijme.2024.101041

Shahsavar, Z., Kafipour, R., Khojasteh, L., & Pakdel, F. (2024). Is artificial intelligence for everyone? Analyzing the role of ChatGPT as a writing assistant for medical students. *Frontiers in Education, 9*, 1457744. https://doi.org/10.3389/feduc.2024.1457744

Song, C., & Song, Y. (2023). Enhancing academic writing skills and motivation: Assessing the efficacy of ChatGPT in AI-assisted language learning for EFL students. *Frontiers in Psychology, 14*, 1260843. https://doi.org/10.3389/fpsyg.2023.1260843

Tafazoli, D. (2024). Exploring the potential of generative AI in democratizing English language education. *Computers and Education Artificial Intelligence, 7*, 100275. https://doi.org/10.1016/j.caeai.2024.100275

Wang, C., Zou, B., Du, Y., & Wang, Z. (2024). The impact of different conversational generative AI chatbots on EFL learners: An analysis of willingness to communicate, foreign language speaking anxiety, and self-perceived communicative competence. *System*, 103533. https://doi.org/10.1016/j.system.2024.103533

Xia, Q., Weng, X., Ouyang, F., Lin, T. J., & Chiu, T. K. (2024). A scoping review on how generative artificial intelligence transforms assessment in higher education. *International Journal of Educational Technology in Higher Education, 21*, 40. https://doi.org/10.1186/s41239-024-00468-z

Yusuf, A., Pervin, N., Román-González, M., & Noor, N. M. (2024). Generative AI in education and research: A systematic mapping review. *Review of Education, 12*(2), e3489. https://doi.org/10.1002/rev3.3489

Zhai, X. (2024). Transforming teachers' roles and agencies in the era of generative AI: Perceptions, acceptance, knowledge, and practices. *Journal of Science Education and Technology*. Advance online publication. https://doi.org/10.1007/s10956-024-10174-0

About the Contributor

Raigul Zheldibayeva Department of Pedagogy and Psychology, Zhetysu University, 040000 Taldykorgan, Kazakhstan.
Email: r.zheldibayeva@zu.edu.kz

Appendix

Appendix I. Prompt for listening practice

Please generate a B1 level (CEFR) story. First, ask me for the topic. If I don't specify a topic, choose one that you deem appropriate for this level. After presenting the story, wait for me to confirm that I have finished listening by typing "I'm done listening" or a similar phrase. Important: Do not show me the cloze passage until I have confirmed I am done listening. To ensure this, after I confirm, send a separate message containing a series of about 10 lines filled with dashes or another symbol to create visual separation. Only after sending this separator message, create a cloze passage based on the story. The cloze passage should have between 5 and 10 blanks. Number the gaps. After I write my responses to the task, please evaluate them. If I get fewer than half of the answers correct, create another story on the same topic and create another new cloze passage. After completing work on the story, ask me for the new topic, or choose the topic yourself. Generate a B1 level conversation on this topic, and then repeat the process described above: wait for my "I'm done listening" confirmation, send the separator, and present a new cloze passage related to this conversation.

https://doi.org/10.22521/edupij.2025.14.51 Published online by Universitepark Press