

Generative AI Agents in Language Learning: A Randomized Field Experiment

Gyeombi Cheon
Korea University Business School
cjsruaql@korea.ac.kr

Yunmin Choi
Korea University Business School
yunminc@korea.ac.kr

Dongwon Lee
Korea University Business School
mislee@korea.ac.kr

Jiye Baik
Korea University Business School
jiyebaik@korea.ac.kr

Abstract

Artificial Intelligence (AI) is revolutionizing education, particularly with advancements in generative AI conversational agents. Our study investigates the effectiveness of these generative AI agents in enhancing English-speaking skills compared to human agents. Through a randomized field experiment involving 363 participants, we found that unrestricted use of AI tools led to a 5.90% improvement in lexical diversity, as measured by the Type-Token Ratio (TTR), highlighting the benefits of self-paced learning. Notably, learners with below-average proficiency experienced a 9.53% improvement in TTR, suggesting AI's potential to bridge educational equity gaps. Moreover, AI tools significantly reduced evaluation apprehension, further enhancing learning outcomes. These findings underscore AI's capacity to provide personalized, anxiety-free learning environments, particularly for students with lower proficiency, and offer valuable insights for integrating AI into educational strategies to foster more inclusive learning experiences.

Keywords: Generative AI, Language Learning, Evaluation Apprehension

1. Introduction

Artificial Intelligence (AI), especially through advancements in Large Language Models (LLMs), is having a profound impact across various fields (Davenport & Mittal, 2022; Kshetri et al., 2023). AI-powered conversational agents, or "AI agents," enhance communication through chatbots and voice-based systems (Sundar, 2020), showing significant potential in education by offering personalized learning tailored to individual needs (Chen et al., 2021). The rise of generative AI has sparked debates about its role as an automated teaching tool and its

potential to replace human educators (Singer, 2024). Given AI's growing influence in education, research has begun to explore its applications across diverse fields (Kazemi, 2023; Sánchez-Ruiz et al., 2023) and to identify effective chatbot design principles (Wu & Yu, 2023).

While previous studies have investigated the impact of AI agents on educational outcomes (e.g., Araujo, 2018), this study specifically examines the role of AI agents as conversation partners in language learning, comparing their effects with those of human partners. In language learning, the choice of conversational partner is crucial. For example, children use more complex language with adults than with peers (Shatz & Gelman, 1973), and foreign language learners' interactions vary based on their conversation partners (Sato, 2007). The integration of AI introduces new dynamics into these interactions, offering advantages like unlimited practice and multimodal features (Shawar, 2017), but also posing challenges such as perceptions of unnatural speech and communication failures (Huang et al., 2022).

Moreover, AI systems often operate as "black boxes," making it difficult to understand how responses are generated, and raising concerns about bias, privacy, and intellectual property (Singer, 2024). These issues are particularly relevant in the context of assessment, originality, and plagiarism (Stokel-Walker, 2022). This underscores the need for empirical research to understand how learners perceive and respond to generative AI agents compared to human partners.

Despite growing interest in AI's role in education, there remains a critical gap in understanding its direct impact when used as a primary learning tool rather than as a supplement. Most research to date has focused on the applicability (Sánchez-Ruiz et al., 2023) and efficiency (Wu & Yu, 2023) of AI in education, particularly as a classroom supplement

(Lee & Jeon, 2022; Ebadi & Amini, 2022; Chien et al., 2022). However, it is essential to empirically assess the effectiveness of AI-driven education as the main mode of learning. As AI takes on a more central role, it could lead to greater student autonomy and a stronger reliance on technology, thereby enhancing personalized learning. Additionally, the shift toward AI as a primary educational tool could necessitate significant changes in the roles of teachers, curriculum design, and assessment methods. Evaluating these potential shifts is crucial, given the transformative potential of generative AI in education. This study, therefore, aims to empirically assess the impact of generative AI as a primary educational tool, offering valuable insights into its viability in real-world learning contexts.

We examine the effectiveness of generative AI in English language learning for several key reasons. First, mastering English—encompassing reading, writing, and speaking—is essential for many, as English proficiency is often a prerequisite for academic and professional opportunities (Wang et al., 2017). However, English remains a barrier due to the high costs and limited accessibility of quality in-person courses (Ruan et al., 2021). Second, proficiency in spoken English highlights significant educational disparities, influenced by factors such as familial and economic backgrounds (Dong, 2023) and exposure to foreign environments. This study focuses on AI agents increasingly used in educational settings, comparing their impact on learning outcomes with that of human conversation partners.

While previous research has not thoroughly explored the effectiveness of AI as an autonomous learning tool outside structured environments, our study compares generative AI agents with human agents in real-world, self-directed learning contexts. This comparison is crucial, as generative AI provides personalized learning experiences that are accessible anytime, anywhere. By examining AI-driven education in these flexible environments, we aim to uncover the true educational potential of generative AI. Additionally, our findings will offer insights into how the growing replacement of human-delivered education with AI influences learning disparities rooted in background proficiencies.

We aim to clarify the effectiveness of generative AI compared to human conversational agents by examining the role of evaluation apprehension. Evaluation apprehension, the fear of negative judgment, can limit both the quantity and quality of communication, thereby hindering language learning (MacIntyre et al., 1997; Li, 2021). However, this apprehension tends to decrease when interacting with

virtual agents, even when conveying the same information (Raveendhran et al., 2020).

This research provides a nuanced understanding of how generative AI facilitates language learning by offering personalized, adaptive experiences that are accessible anytime, anywhere. This potential to democratize education could help reduce disparities in language proficiency. Additionally, we aim to explore how the shift from human-delivered to AI-driven education might influence existing disparities in learning outcomes, particularly those based on students' prior proficiency levels. Understanding this impact is crucial, as it will reveal whether AI can equalize educational opportunities or if it might inadvertently widen the gap for those already at a disadvantage. To investigate the transformative impact of generative AI as a substitute for in-person learning, especially in addressing educational inequalities, we focus on the following research question: How do AI conversational agents affect learners' language-speaking performance?

To answer our research question, we partnered with a South Korean company that developed a mobile application offering English-speaking sessions via generative AI virtual agents. We conducted a randomized field experiment over four weeks, assigning participants to groups using AI agents with unlimited access, AI agents with time restrictions, or traditional learning methods.

Our empirical analyses demonstrate that generative AI conversational agents significantly enhance English-speaking skills, particularly in lexical diversity, as measured by the Type-Token Ratio (TTR). Unrestricted use of AI learning tools led to a 5.90% improvement in TTR, highlighting the benefits of self-paced learning. Learners with below-average proficiency saw the most significant gains, with a 9.53% improvement in TTR, suggesting that generative AI can bridge educational equity gaps by offering personalized support to those who need it most. Further analysis revealed that generative AI tools effectively engage students who might otherwise struggle with traditional methods. These tools were particularly beneficial for learners with lower proficiency, indicating that AI can support educational equity by providing more tailored learning experiences. Moreover, generative AI agents significantly reduce evaluation apprehension (EA), the fear of being negatively judged, contributing to better learning outcomes. Mediation analysis suggested that reducing EA through AI tools could lead to a 1.162% improvement in TTR.

Our research contributes three key aspects to the literature on learning with generative AI agents and learning performance. First, by comparing AI agents

under different time constraints, we explore their potential for diverse educational strategies and their application as primary tools in self-directed learning environments. Second, we investigate the impact of AI agents on learners with varying background proficiencies, showing that AI can mitigate educational disparities, particularly in foreign language acquisition. Third, we examine evaluation apprehension, explaining why AI and human conversation partners yield different outcomes in language learning. Our study reveals that AI agents can reduce learners' anxiety about being evaluated, contributing to a more comfortable and effective learning experience.

These findings have practical implications for educators and policymakers, suggesting that integrating AI into educational strategies can enhance personalized and flexible learning, especially for students who struggle with traditional methods.

2. Literature Review

2.1. Generative AI in Educational Context

Generative AI, a subfield of AI that generates content like text, images, audio, and video (McKinsey, 2023), has quickly become integrated into education. AI has shown promise across various fields, from enhancing computational thinking in novice programmers (Kazemi, 2023) to improving mathematics (Sánchez-Ruiz et al., 2023) and advancing medical education (Oh et al., 2023). Beyond these applications, research is exploring how AI can improve learner experiences. For instance, Wu and Yu (2023) found that chatbot designs with human-like avatars, gamification, and emotional intelligence enhance learning outcomes. Similarly, Lim et al. (2023) highlighted the importance of user-centered design in AI tools, focusing on learners' perceptions, satisfaction, and engagement with ChatGPT.

2.2. Effectiveness of Generative AI in Language Learning

Research increasingly supports the positive effects of AI on language acquisition (Tai, 2022; Timpe-Laughlin et al., 2022; Xu et al., 2022), but the impact of different educational strategies using the same AI platform is less explored (Lin & Mubarak, 2021). While some studies compare different types of AI agents (Chien et al., 2022; Yang et al., 2022), few investigate varying strategies with the same AI tools. Our study examines the impact of time constraints within a generative AI learning context, focusing on

the benefits of self-paced learning, where learners set their own speed and schedule. This flexibility contrasts with traditional methods that rely heavily on continuous teacher guidance (García Botero et al., 2019; Stockwell & Reinders, 2019). By comparing learners' performance under fixed versus unlimited time with AI agents, we aim to understand how AI-driven self-paced learning compares to more structured environments.

2.3. Learners' Background Proficiency

Existing research has explored the effectiveness of generative AI agents, but it has not sufficiently addressed how individual learner characteristics, such as background proficiency, influence outcomes (Jeon et al., 2023). This study examines how learners' prior proficiency affects the efficacy of generative AI in language learning. Previous studies indicate that learning interventions are more effective when tailored to students' prior knowledge: novices benefit more from examples, while experienced learners excel with problem-solving (Koedinger et al., 2012). Additionally, prior knowledge reduces cognitive load, leading to better learning outcomes (van Riesen et al., 2022). Students with more background knowledge engage more effectively, while those with less are less inclined to seek instructional support (Dong et al., 2020). These findings suggest that generative AI's effectiveness may vary based on learners' proficiency levels. By exploring how AI impacts learners with different levels of prior proficiency, this study aims to determine whether AI can offer tailored support to address these differences, potentially reducing educational disparities.

2.4. Evaluation Apprehension

To understand the differential impact of human and AI agents on learning performance, a theoretical framework is necessary. Evaluation Apprehension, the fear of being judged, can hinder learning by distracting attention and triggering performance anxiety (Cottrell et al., 1968; Schlenker & Leary, 1982). This anxiety is typically less when interacting with AI, as people feel less judged by machines (Siemon, 2023). We apply this concept to educational environments utilizing generative AI to explain its effectiveness. Reducing this anxiety can create a more comfortable learning environment, potentially leading to more effective language learning.

3. Research Methodology

3.1. Experimental Background

We partnered with a South Korean company offering a mobile application for English-speaking sessions using a generative AI-based virtual agent. This AI engages users in verbal conversations tailored to their interests and previous interactions. After installing the app, learners can start conversations by selecting the AI agent. Like other AI-driven speaking applications, this app provides real-time feedback on incorrect or unclear utterances, offering grammatical corrections and alternative expressions for each, along with comprehensive feedback after the conversation ends.

The control group practiced English speaking with human tutors via mobile devices, engaging in voice sessions with native English instructors through a South Korean mobile service. Thus, both the treatment and control groups conducted their English learning via mobile devices—one interacting with the AI agent within the app, and the other participating in voice calls with human tutors.

3.2. Dependent Variables

We assess the outcome of English learning through various methods. First, we evaluate the linguistic complexity of spoken content using the Type-Token Ratio (TTR)¹, a widely adopted measure for assessing lexical diversity (Biber, 2007). TTR indicates the variety of vocabulary used. Higher TTR values denote greater lexical diversity and complexity. To obtain the TTR scores of participants before and after the learning sessions, we conducted two mock TOEIC² Speaking tests before and after the four-week learning period. This method allows us to objectively quantify changes in participants' vocabulary use and language complexity as a result of the intervention.

We also measured perceived student engagement through a post-experiment survey based on items from Chiu (2021). The survey used a five-point Likert scale to evaluate cognitive engagement with questions such as: 1) "When learning through generative AI/human, I sought clarification by looking up words or asking the conversation agent again." 2) "I applied the English expressions learned through generative AI/human to conversations." 3) "I related the content I was learning to what I had previously learned."

¹ Type-Token Ratio (TTR) is computed using the formula: (Number of Lexical Items / Total Number of Words) * 100 (Biber, 2007)

3.3. Experiment Design

Our study aims to assess the educational impact of generative AI as the primary learning tool. We recruited participants from various online university student communities in Korea. Participants were briefed on the procedures and randomly assigned to one of three groups: (1) a human conversational agent group with sessions three times a week for 20 minutes, (2) a restricted AI group with the same session frequency and duration, and (3) an unrestricted AI group with no study time limits. Each group consisted of 121 participants, and the English-speaking program lasted four weeks.

Before the sessions began, participants took a pretest—a mock TOEIC Speaking Test provided by ETS, commonly used by South Korean undergraduates. Responses were recorded, transcribed using a commercial speech-to-text algorithm, and analyzed for lexical diversity using the Type-Token Ratio (TTR). After the four-week learning period, participants took a post-test with different TOEIC Speaking questions. Additionally, surveys were conducted post-experiment to assess perceived engagement and evaluation apprehension.

Participants were required to complete a survey to collect background information, including gender, age, education, and recent TOEIC scores. The pre-experiment survey gathered data on gender, age, and student status (undergraduate or graduate), while the post-experiment survey inquired about additional study materials used, preview function usage, and prior experience with the learning tool.

Additionally, we gather statistics on each participant's degree of participation by recording the average duration of attended sessions and the intervals between sessions. These values are incorporated into our final dataset. To ensure data quality and consistency, we remove outliers and exclude participants who dropped out during the experiment. After applying these criteria, we have 248 participants: 99 in the control group, 72 in the restricted AI treatment group, and 77 in the unrestricted AI treatment group.

3.4. Empirical Model

To evaluate the effectiveness of generative AI in learning English speaking, we developed a model to measure changes in performance. Specifically, we

² TOEIC stands for "Test of English for International Communication."

modeled the treatment effect of AI agents versus human agents using the following regression equation:

$$LexicalDiversity_i = \beta_0 + \beta_1 GenAI_i + \beta_2 Control_i + \varepsilon_i \quad (1)$$

In this model, the dependent variable, *LexicalDiversity_i*, represents the percentage change in Type-Token Ratio (TTR), which measures the enhancement in lexical diversity. The independent variables include the treatment group indicator *GenAI_i*, which denotes whether participants used AI conversational agents or human agents, and a vector of control variables *Control_i*. These control variables comprise gender, other demographic factors, the count of sessions attended, the duration of sessions, the average interval between sessions, age, and whether participants previewed the study materials. We provide a comprehensive list of these variables in Table 1 and present descriptive statistics of the variables across the three groups in Table 2.

Variables	Definition
<i>LexicalDiversity</i>	Lexical diversity of spoken content using the TTR (Type-Token Ratio). Calculated as the percentage change in TTR from pre-test to post-test by comparing the post-test TTR to the pre-test TTR, dividing the difference by the pre-test TTR, and multiplying by 100.
<i>Student Engagement</i>	The mean score of perceived engagement obtained from the post-experiment survey
<i>Gen AI</i>	Binary variable where 1 represents the use of generative AI and 0 represents human agents
<i>Other Materials</i>	Binary variable indicating whether the participant used other learning tools concurrently
<i>Preview</i>	Binary variable indicating whether the participant used the preview function
<i>Age</i>	Participant's age
<i>Gender</i>	Binary variable indicating the gender of the participant
<i>Duration</i>	Total learning time of the participant (in seconds)
<i>Interval</i>	Average learning interval of the participant (in days)
<i>Count</i>	Number of learning sessions attended by the participant
<i>Level</i>	Participants' self-reported English proficiency level before the experiment (TOEIC score 0~990)

Table 1. Variable Definitions.

Group	Variables	Mean	Median	Minimum	Maximum
Control Group	<i>Other Materials</i>	0.247	0	0	1
	<i>Preview</i>	0.618	1	0	1
	<i>Age</i>	25.48	24	20	45
	<i>Gender</i>	0.707	1	0	1
	<i>Duration</i>	12,667	13,200	3,600	15,600
	<i>Interval</i>	3.417	2.818	2	18.5
	<i>Count</i>	10.56	11	3	13
	<i>Level</i>	834.8	850	515	990
	Treatment Group1	<i>Other Materials</i>	0.166	0	0
<i>Preview</i>		0.606	1	0	1
<i>Age</i>		25.27	24	19	44

Treatment Group2	<i>Gender</i>	0.833	1	0	1
	<i>Duration</i>	6,226	4,232	54	23,622
	<i>Interval</i>	5.396	3.5	1	19
	<i>Count</i>	6.194	4.5	2	14
	<i>Level</i>	842.3	865	440	990
	<i>Other Materials</i>	0.323	0	0	1
	<i>Preview</i>	0.602	1	0	1
	<i>Age</i>	26.66	24	20	49
	<i>Gender</i>	0.662	1	0	1
	<i>Duration</i>	5,914	4,489	25,676	56
	<i>Interval</i>	4.632	3.5	1	15.5
	<i>Count</i>	6.286	5	2	23
	<i>Level</i>	843.8	850	426	985

Table 2. Descriptive Statistics.

4. Results

Our main analysis compared improvements in lexical diversity, measured by the Type-Token Ratio (TTR), between groups learning with generative AI and those with human agents. The results are presented in Table 3, column (1). The findings indicate that AI-based learning agents have a positive coefficient, suggesting a more favorable impact on TTR improvement compared to human-led sessions. Specifically, participants using AI agents experienced, on average, a 4.26% greater improvement in TTR compared to those using human agents.

To examine the impact of different time restrictions within the AI groups, we analyze the results separately for the restricted and unrestricted AI groups. As shown in Table 3, column (2), the unrestricted AI group (Treatment Group 2) demonstrates a significant positive effect on TTR improvement. Participants in this group showed an average improvement of 5.90% in TTR compared to the control group. In contrast, the restricted AI group (Treatment Group 1) showed only a 2.52% increase in TTR, which was not statistically significant. These findings suggest that unrestricted access to AI agents may lead to better learning outcomes compared to restricted usage.

This suggests that the independent use of generative AI could have an educational impact similar to that of human tutors. In contrast to earlier studies, which have demonstrated that AI can either effectively improve speaking skills (Kim, 2021) or have minimal influence on academic performance (Deveci Topal et al., 2021), these results indicate that generative AI can significantly enhance overall language proficiency, comparable to the effect of human instructors.

Our findings also underscore the importance of learning autonomy when using AI-based agents. The unrestricted AI group, which allowed participants to study at their own pace, showed more significant

improvements in TTR. This highlights the potential of generative AI to maximize learning outcomes through self-paced learning conditions, reflecting a realistic and flexible learning scenario. In contrast, the restricted AI group's performance was comparable to the human-led sessions, suggesting that time constraints may limit the benefits of AI agents. Therefore, AI-based learning interventions should allow flexibility to fully enhance their effectiveness. These results emphasize the critical role of self-paced learning conditions in maximizing the benefits of AI in education. By allowing learners to control the pace and timing of their studies, AI-based agents can better accommodate individual needs and learning styles, thereby enhancing overall educational outcomes.

	(1)	(2)
	Lexical Diversity	
<i>Intercept</i>	-12.157* (6.234)	-11.895* (6.219)
<i>Gen AI</i>	4.258* (2.305)	-
<i>Gen AI(T1)</i>	-	2.522 (2.613)
<i>Gen AI(T2)</i>	-	5.902** (2.582)
<i>Others</i>	3.591 (2.178)	3.270 (2.184)
<i>Preview</i>	-0.395 (1.901)	-0.459 (1.896)
<i>Age</i>	0.343* (0.180)	0.320* (0.180)
<i>Gender</i>	-0.780 (2.112)	-0.509 (2.115)
<i>Duration</i>	0.001 (0.001)	0.001 (0.001)
<i>Interval</i>	0.104 (0.315)	0.163 (0.317)
<i>Count</i>	0.377 (0.570)	0.351 (0.569)
<i>Adjusted R²</i>	0.024	0.030
<i>N</i>	181	

Table 3. Treatment Effects on English Skills.

Notes. The sample size (N) is 181 because, out of the initial 248 participants, those who did not respond to the survey were excluded (248 -> 224). Subsequently, participants who did not complete the post-test were also excluded (224 -> 181). Standard errors in parentheses.

***p < 0.01; **p < 0.05; *p < 0.1.

5. Heterogeneous Effects

Background proficiency may significantly influence learning with generative AI agents. To investigate this, we collected participants' official English language scores (TOEIC) before the study. These scores allowed us to classify participants into above-average and below-average proficiency levels. Participants with scores above the median were placed

in the "above group," and those with scores below the median in the "below group." A variable called "below_avg" indicated each participant's proficiency level. By segmenting participants based on their background proficiency, we aimed to determine which group benefits more from generative AI agents. Specifically, if generative AI agents prove more effective for learners with lower proficiency, they could help bridge educational equity gaps.

Table 5 presents the results of our analysis, showing whether AI agents facilitate better learning than human agents for participants with lower background proficiency. The results reveal a significant interaction effect between the unrestricted AI group and below-average proficiency learners. Specifically, in the unrestricted AI group (Treatment Group 2), participants with below-average proficiency exhibited a significant 9.53% improvement in their Type-Token Ratio (TTR). This suggests that the generative AI agent is particularly beneficial for learners with lower English proficiency levels. Participants with lower foundational English skills showed greater engagement and improvement when using the generative AI agent compared to other groups. These findings underscore the importance of providing a flexible learning environment, especially for learners with lower proficiency.

Previous studies have indicated that AI technologies do not benefit all students equally, with only highly motivated or high-achieving students typically reaping the most rewards, while digital tools may be less effective for those with very high or low prior abilities (Bonneton-Botté et al., 2020). Our findings build on this existing research by demonstrating that prior ability can result in diverse outcomes in language performance.

	(1)	(2)
	Lexical Diversity	
<i>Intercept</i>	-17.034** (7.194)	-16.246* (7.039)
<i>Gen AI</i>	2.577 (3.195)	-
<i>Gen AI(T1)</i>	-	3.350 (3.560)
<i>Gen AI(T2)</i>	-	2.107 (3.448)
<i>Gen AI*Below</i>	4.488 (3.902)	-
<i>Gen AI(T1)*Below</i>		-1.643 (4.573)
<i>Gen AI(T2)*Below</i>		9.536** (4.378)
<i>Below</i>	-4.546 (3.081)	-4.565 (3.011)
<i>Others</i>	3.716 (2.270)	3.047 (2.233)
<i>Preview</i>	1.776	1.362

	(1.954)	(1.916)
Age	0.441**	0.417**
	(0.188)	(0.185)
Gender	0.906	1.233
	(2.245)	(2.200)
Duration	-0.001	-0.001
	(0.001)	(0.001)
Interval	0.245	0.311
	(0.317)	(0.315)
Count	0.785	0.504
	(0.609)	(0.603)
R ²	0.137	0.188
N	145	

Table 4. Heterogeneity in Treatment Effects.

Notes. The sample size (N) is 141 because participants without TOEIC English scores were also excluded. Standard errors in parentheses.

***p < 0.01; **p < 0.05; *p < 0.1.

6. Mechanism and Robustness

6.1. Exploration of Potential Mechanisms

Evaluation apprehension, the fear of being negatively judged, can significantly hinder communication quality and quantity (MacIntyre et al., 1997) and negatively affect language learning outcomes (Li, 2021). However, research suggests that interactions with virtual agents can mitigate this apprehension, even with identical content (Raveendhran et al., 2020). Based on these insights, we hypothesize that generative AI conversational agents can significantly reduce evaluation apprehension, thereby enhancing learning outcomes.

To further investigate the mechanism, we conducted a causal mediation analysis to determine how much of the effect of AI agents on TTR improvement is mediated through evaluation apprehension. The Average Causal Mediation Effect (ACME) was 1.162, suggesting that AI agents contribute to a 1.162% improvement in TTR by reducing evaluation apprehension. This finding provides valuable insights into the potential of AI agents to create a comfortable learning environment that mitigates anxiety and enhances learning outcomes. However, while the p-value is statistically significant, it is important to note that the confidence interval includes zero, indicating that this mechanism is plausible but not definitive. We therefore acknowledge evaluation apprehension as a potential mechanism, but with caution regarding its certainty.

	Coefficient	95% CI Lower	95% CI Upper
ACME (Average Causal	1.162*	-0.196	2.620

Mediation Effect			
ADE (Average Direct Effect)	3.097	-2.631	8.823
Total Effect	4.259*	-0.951	9.40
Proportion Mediated	0.273	-0.995	3.24

Table 5. Mediation Effects.

Notes. Simulations: 1000

***p < 0.01; **p < 0.05; *p < 0.1.

6.2. Further Analysis

To ensure the robustness of our findings, we conducted additional analyses using "Student Engagement," the mean score of perceived engagement from the post-experiment survey, as the dependent variable. The results were consistent, reinforcing the reliability of our findings.

Cognitive Student Engagement (1)	
Constant	5.663** (2.088)
GenAI(T1) *	1.058 (0.647)
Below	1.467* (0.603)
Other Controls	Yes
R ²	0.185
N	166

Table 6. Further Analysis.

Notes. . +p<0.1, *p<0.05, **p<0.01, ***p<0.001

To ensure the robustness and generalizability of our findings, we replicated our results using the Hapax Dislegomena Ratio (HDR) instead of the Type-Token Ratio (TTR). HDR measures the ratio of words that occur exactly twice in a text to the total number of words, providing another perspective on lexical diversity. This metric characterizes a text independently of its length (Garcia & Martin, 2006). For vocabulary richness, HDR is a reliable indicator (Türkoğlu et al., 2007)

	(1)	(2)
Lexical Diversity (HDR)		
Intercept	3.274 (22.362)	4.136 (22.331)
Gen AI	12.513 (8.268)	-
Gen AI(T1)	-	6.793 (9.381)
Gen AI(T2)	-	17.928* (9.271)
Others	-10.650 (7.814)	-11.708 (7.843)
Preview	-0.797	-1.009

	(6.820)	(6.810)
<i>Age</i>	0.187	0.112
	(0.648)	(0.649)
<i>Gender</i>	-8.320	-7.426
	(7.577)	(7.595)
<i>Duration</i>	0.001	0.001
	(0.001)	(0.001)
<i>Interval</i>	1.560	1.754
	(1.131)	(1.139)
<i>Count</i>	-0.873	-0.959
	(2.046)	(2.043)
<i>Adjusted R²</i>	0.057	0.066
<i>N</i>	181	

Table 7. Further Analysis.

Notes. Standard errors in parentheses.

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$.

Table 7 presents the regression results focusing on HDR as the dependent variable. The results indicate that the unrestricted AI group (Gen AI (T2)) shows a significant improvement in HDR, with a 17.928% increase compared to the control group. This significant coefficient underscores the effectiveness of AI agents in enhancing lexical diversity through HDR, further validating our findings from the TTR metric. These findings consistently highlight the potential of generative AI to foster substantial improvements in language learning outcomes across different measures of lexical diversity.

7. Discussion and Future Research

Our study highlights the effectiveness of generative AI conversational agents in improving language proficiency, particularly in enhancing lexical diversity as measured by the Type-Token Ratio (TTR). The key findings are summarized as follows: Generative AI agents improve TTR by 4.26%, indicating that these tools are at least as effective as human agents in enhancing language proficiency. Participants with unrestricted access to AI agents (Treatment Group 2) experienced a 5.90% improvement in TTR, underscoring the importance of allowing learners to study at their own pace, which can lead to better educational outcomes. Learners with below-average proficiency benefit the most from AI agents, showing a 9.53% improvement in TTR. This suggests that generative AI can play a crucial role in bridging educational equity gaps by providing more effective support to learners who need it the most. Generative AI agents help reduce evaluation apprehension, the fear of being negatively judged, which contributes to better learning outcomes. Mediation analysis revealed that 27.3% of the AI agents' impact on TTR improvement is due to reduced evaluation apprehension, highlighting the importance

of creating comfortable learning environments. The robustness of these findings was confirmed through additional analyses using survey data on student engagement and an alternative lexical diversity metric, the Hapax Dislegomena Ratio (HDR). These analyses consistently show that AI agents enhance learning outcomes by reducing anxiety and promoting engagement. In summary, generative AI agents significantly enhance learning outcomes by providing personalized and flexible learning environments that reduce anxiety and promote engagement. This is especially beneficial for learners with lower proficiency, supporting educational equity and highlighting the potential of AI-based educational tools.

Our research contributes to the educational field related to generative AI in several significant aspects. Firstly, our findings strengthen the evidence regarding the effectiveness of various approaches in generative AI by comparing AI agents under different time constraints. This suggests that generative AI learning can be effectively utilized for self-directed learning, supporting the potential of AI to replace human agents when learners engage autonomously outside the classroom. Secondly, we add to the existing literature on the effects of educational interventions based on students' prior proficiency levels. Previous research has indicated that the effectiveness of learning tools varies depending on students' prior knowledge, with novice learners benefiting from examples and model answers, while more experienced learners benefit from problem-solving activities (Koedinger et al., 2012). Our findings extend this research by showing that learners with lower background proficiency benefit more from generative AI. Thirdly, we introduce a new theoretical mechanism by leveraging evaluation apprehension as a psychological factor to explain the differences in learning outcomes between AI and human conversation partners. Consequently, our findings contribute to the literature on comparative studies between generative AI and human agents by providing insights into why AI can foster a more comfortable and effective learning environment.

Our study presents significant practical implications for educational institutions and policymakers by demonstrating the potential of generative AI conversational agents to enhance language learning. Notably, our findings emphasize the particular benefits for learners with lower English proficiency levels, who exhibited substantial improvements when using these AI tools. This suggests that integrating generative AI into educational frameworks can effectively support students with varying language abilities and help bridge educational equity gaps. Moreover, our results

underscore the importance of the self-paced nature of AI-based learning tools. The flexibility provided by these tools allows learners to study at their own pace, which is especially beneficial for those who might struggle in traditional, time-constrained environments.

However, this study has limitations that present opportunities for future research. First, future studies could explore the long-term effects of generative AI, as our one-month learning period may have been influenced by external factors such as offline discussions with peers. Second, the exclusive focus on the Korean population limits the generalizability of our findings to other cultural or linguistic contexts. Future research should aim to replicate this study in different countries and with diverse populations to determine whether the observed effects hold true across various cultural settings. Third, it is an interesting observation that some participants dropped out of the study, even though they voluntarily participated, and the experiment lasted only one month. Notably, the dropout rate was higher in the group that learned with generative AI compared to the group that learned with human tutors. Exploring how this higher dropout rate might have influenced the results could be an intriguing topic for future research. Forth, while our methodology offers valuable insights into learning outcomes, it primarily focuses on particular indicators of English proficiency, such as lexical diversity. This approach may not capture the full range of language performance. Future research could be enhanced by incorporating evaluations from native speakers, providing a more detailed analysis and a broader understanding of the effects of generative AI on language acquisition. Lastly, future research could conduct additional field experiments or surveys to further investigate the reasons and mechanisms behind the language learning outcomes associated with generative AI.

8. References

- Araujo, T. (2018). Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. *Computers in human behavior*, 85, 183-189.
- Swales, J. M. (2009). Discourse on the move: Using corpus analysis to describe discourse structure. *Language*, 85(3), 694-696..
- Bonneton-Botté, N., Fleury, S., Girard, N., Le Magadou, M., Cherbonnier, A., Renault, M., & Jamet, E. (2020). Can tablet apps support the learning of handwriting? An investigation of learning outcomes in kindergarten classroom. *Computers & Education*, 151, 103831.
- Chen, X., Zou, D., Cheng, G., & Xie, H. (2021). "Artificial intelligence-assisted personalized language learning: systematic review and co-citation analysis". In *2021 International Conference on Advanced Learning Technologies (ICALT)* (pp. 241-245). IEEE.
- Chien, Y. C., Wu, T. T., Lai, C. H., & Huang, Y. M. (2022). "Investigation of the influence of artificial intelligence markup language-based LINE ChatBot in contextual English learning" *Frontiers in Psychology*, 13, 785752.
- Chiu, T. K. (2021). "Digital Support for Student Engagement in Blended Learning Based on Self-Determination Theory," *Computers in Human Behavior*, 124, 106909.
- Cottrell, N. B., Wack, D. L., Sekerak, G. J., & Rittle, R. H. (1968). Social facilitation of dominant responses by the presence of an audience and the mere presence of others, *Journal of personality and social psychology*, 9(3), 245.
- Davenport, T. H., & Mittal, N. (2022). How Generative AI Is Changing Creative Work, *Harvard Business Review*.
- Deveci Topal, A., Dilek Eren, C., & Kolburan Geçer, A. (2021). Chatbot application in a 5th grade science course. *Education and Information Technologies*, 26(5), 6241-6265.
- Dong, L. (2024). Examining the relationship between socioeconomic status, self-regulated learning strategies, and writing proficiency in English as a second language learning context. *Journal of Educational Psychology*, 116(5), 686.
- Ebadi, S., & Amini, A. (2022). Examining the roles of social presence and human-likeness on Iranian EFL learners' motivation using artificial intelligence technology: A case of CSIEC chatbot. *Interactive Learning Environments*, 32(2), 655-673.
- García Botero, G., Questier, F., & Zhu, C. (2019). Self-directed language learning in a mobile-assisted, out-of-class context: do students walk the talk? *Computer Assisted Language Learning*, 32(1-2), 71-97.
- Garcia, A. M., & Martin, J. C. (2006). Function words in authorship attribution studies. *Literary and Linguistic Computing*, 22(1), 49-66.
- Huang, W., Hew, K. F., & Fryer, L. K. (2022). Chatbots for language learning—Are they really useful? A systematic review of chatbot-supported language learning. *Journal of Computer Assisted Learning*, 38(1), 237-257.
- Jeon, J., Lee, S., & Choi, S. (2023). A systematic review of research on speech-recognition chatbots for language learning: Implications for future directions in the era of large language models. *Interactive Learning Environments*, 1-19.
- Kazemi, R. (2023). Artificial intelligence techniques in advanced concrete technology: A comprehensive survey on 10 years research trend, *Engineering Reports*, 5(9), e12676.
- Kim, H. S. (2021). Is it beneficial to use AI chatbots to improve learners' speaking performance?. *Journal of Asia TEFL*, 18(1), 161-178.
- Koedinger, K. R., Corbett, A. T., & Perfetti, C. (2012). The Knowledge-Learning-Instruction framework: Bridging the science-practice chasm to enhance robust student learning, *Cognitive science*, 36(5), 757-798.

- Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Koohang, A., & Wright, R. (2023). So what if ChatGPT wrote it? Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, 71, 102642.
- Lee, S., & Jeon, J. (2022). Visualizing a disembodied agent: Young EFL learners' perceptions of voice-controlled conversational agents as language partners. *Computer Assisted Language Learning*, 37(5-6), 1048-1073.
- Li, X. (2021). EFL teachers' apprehension and L2 students' classroom engagement, *Frontiers in Psychology*, 12, 758629.
- 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. *The international journal of management education*, 21(2), 100790.
- Lin, C. J., & Mubarak, H. (2021). Learning analytics for investigating the mind map-guided AI chatbot approach in an EFL flipped speaking classroom, *Educational Technology & Society*, 24(4), 16-35.
- MacIntyre, P. D., Noels, K. A., & Clément, R. (1997). Biases in self-ratings of second language proficiency: The role of language anxiety, *Language learning*, 47(2), 265-287
- McKinsey. (2023). What is ChatGPT, DALL-E, and generative AI? <https://www.mckinsey.com/featuredinsights/mckinsey-explainers/what-is-generativeai>
- Oh, N., Choi, G. S., & Lee, W. Y. (2023). ChatGPT goes to the operating room: evaluating GPT-4 performance and its potential in surgical education and training in the era of large language models, *Annals of Surgical Treatment and Research*, 104(5), 269.
- Raveendhran, R., Fast, N. J., & Carnevale, P. J. (2020). Virtual (Freedom From) Reality: Evaluation Apprehension and Leaders' Preference for Communicating Through Avatars, *Computers in Human Behavior*, 111, 106415
- Ruan, S., Jiang, L., Xu, Q., Liu, Z., Davis, G. M., Brunskill, E., & Landay, J. A. (2021). Englishbot: An ai-powered conversational system for second language learning. *In 26th international conference on intelligent user interfaces* (pp. 434-444).
- Sánchez-Ruiz, L. M., Moll-López, S., Nuñez-Pérez, A., Moraño-Fernández, J. A., & Vega-Fleitas, E. (2023). ChatGPT challenges blended learning methodologies in engineering education: A case study in mathematics, *Applied Sciences*, 13(10), 6039.
- Sato, M. (2007). Social relationships in conversational interaction: A comparison of learner-learner and learner-NS dyads. *JALT Journal*, 29(2), 183.
- Schlenker, B. R., & Leary, M. R. (1982). Social anxiety and self-presentation: A conceptualization model, *Psychological bulletin*, 92(3), 641.
- Shatz, M., & Gelman, R. (1973). The development of communications skills: Modifications in the speech of young children as a function of listener. *Monographs of the Society for Research in Child Development*, 38, 1-37.
- Shawar, B. A. (2017). Integrating CALL systems with chatbots as conversational partners. *Computación y Sistemas*, 21(4), 615-626.
- Siemon, D. (2023). Let the computer evaluate your idea: evaluation apprehension in human-computer collaboration, *Behaviour & Information Technology*
- Singer, N. (2024). Will chatbots teach your children? *New York Times* <https://www.proquest.com/newspapers/will-chatbots-teach-your-children/docview/2915463463/se-2>
- Stockwell, G., & Reinders, H. (2019). Technology, motivation and autonomy, and teacher psychology in language learning: Exploring the myths and possibilities, *Annual Review of Applied Linguistics*, 39, 40-51.
- Stokel-Walker, C. (2022). AI bot ChatGPT writes smart essays—should professors worry?. *Nature*.
- Sundar, S. S. (2020). Rise of machine agency: A framework for studying the psychology of human-AI interaction (HAI). *Journal of Computer-Mediated Communication*, 25(1), 74-88.
- Tai, T. Y., & Chen, H. H. J. (2022). The impact of intelligent personal assistants on adolescent EFL learners' listening comprehension, *Computer Assisted Language Learning*, 1-28.
- Timpe-Laughlin, V., Sydorenko, T., & Daurio, P. (2022). Using spoken dialogue technology for L2 speaking practice: What do teachers think? *Computer Assisted Language Learning*, 35(5-6), 1194-1217.
- Türkoğlu, F., Diri, B., & Amasyalı, M. F. (2007). Author attribution of Turkish texts by feature mining. *In Advanced intelligent computing theories and applications: With aspects of theoretical and methodological issues. Third International Conference on Intelligent Computing (ICIC 2007), Qingdao, China, August 21-24, 2007, Proceedings 3* (pp. 1086-1093). Springer Berlin Heidelberg.
- van Riesen, S. A., Gijlers, H., Anjewierden, A. A., & de Jong, T. (2022). The influence of prior knowledge on the effectiveness of guided experiment design, *Interactive Learning Environments*, 30(1), 17-33.
- Wang, H., Smyth, R., & Cheng, Z. (2017). The economic returns to proficiency in English in China. *China Economic Review*, 43, 91-104.
- Wu, R., & Yu, Z. (2024). Do AI chatbots improve students learning outcomes? Evidence from a meta-analysis, *British Journal of Educational Technology*, 55(1), 10-33
- Xu, Y., Aubele, J., Vigil, V., Bustamante, A. S., Kim, Y. S., & Warschauer, M. (2022). Dialogue with a conversational agent promotes children's story comprehension via enhancing engagement, *Child Development*, 93(2), e149-e167.
- Yang, C. T. Y., Lai, S. L., & Chen, H. H. J. (2022). The impact of intelligent personal assistants on learners' autonomous learning of second language listening and speaking, *Interactive Learning Environments*, 1-21.