



This work is licensed
under a Creative
Commons Attribution
4.0 International
License.

Beyond perception: Longitudinal insights into chatbot-assisted EFL learning

Steven G.B. MacWhinnie

Hirosaki Gakuin University, Japan

smacwhinnie@gmail.com

While AI chatbots are increasingly integrated into English as a Foreign Language (EFL) instruction, many prior studies emphasize learner perceptions rather than objective linguistic development. This longitudinal study analyzed weekly chatbot interactions from first-year English majors in Japan ($n = 18$) over two academic terms. Using chat logs analyzed with paired t-tests, and ANOVAs, this study tracked lexical diversity, in terms of type token ratio (TTR), mean sentence length, expanded answers, self-correction, total turns, and follow-up questions. Results showed initial gains but uneven patterns across terms. In Spring, students significantly increased their total turns ($t(17) = 2.42, p = .028, d^2 = 0.57$) and demonstrated a steady rise in self-correction across sessions ($F(10, 170) = 3.34, p < .001, \eta_p^2 = .16$), while sentence length and expansion trended upward without reaching significance. In Fall, lexical diversity increased significantly ($t(17) = 5.51, p < .001, d^2 = 1.30$), and follow-up questions showed a marginal increase ($t(17) = 2.17, p = .055, d^2 = 0.51$), while expanded answers declined. Overall, the data suggested recurring trends across semesters, though the specific areas of growth differed. These findings indicate that chatbots can foster early improvements in sentence elaboration, interaction, and self-monitoring, but vocabulary and expansion plateau without varied pedagogical support. Implications include diversifying chatbot tasks, setting explicit lexical goals, and integrating teacher mediation to sustain growth.

Keywords: artificial intelligence, language learning, ChatGPT, smartphone applications, lexical analysis

Introduction

The integration of artificial intelligence (AI) into language education has accelerated in recent years. For English as a Foreign Language (EFL) learners in Japan, where daily chances to engage in extended English conversation are limited, chatbots can offer accessible and low-pressure interaction outside of class. They can supplement traditional instruction by providing immediate responses, consistent

availability, and a flexible environment for practicing communicative skills (Ji et al., 2022; Labadze et al., 2023).

Much of the published research on AI chatbots in language learning, however, has emphasized *learner perceptions and technology acceptance* rather than demonstrable linguistic gains (Chan & Hu, 2023; MacWhinnie, 2025; Polakova & Klimova, 2024). Some research has reported moderate benefits for speaking and writing when chatbots are integrated into EFL settings but also stress the lack of longitudinal studies that track actual language improvement over time (Ballidağ & Aydin, 2025; Polakova & Klimova, 2024). Reported challenges include limited vocabulary development, authenticity of interaction, and the risk of plateauing when activities remain repetitive (Antony & Ramnath, 2023; Viktorivna et al., 2022). As a result, researchers have called for empirical studies that examine *how students' language changes across extended chatbot use*, beyond initial novelty effects (MacWhinnie, 2025; Zou et al., 2023).

The present study addresses this gap by analyzing chatbot interactions from first-year English majors at a Japanese university over two academic terms. Building on earlier work that investigated students' perceptions of audio-enabled chatbot use (MacWhinnie, 2025), this classroom-based exploratory study focuses on objective measures of linguistic development. Specifically, it tracks changes in lexical diversity, sentence length, expanded answers, self-correction, follow-up questions, and total turns across weekly chatbot sessions.

By combining paired *t*-tests and ANOVAs on interaction logs, this research highlights not only areas of improvement but also points where progress plateaued. In doing so, it contributes empirical evidence to ongoing debates about the role of AI chatbots in language learning and provides insights for instructors seeking to integrate such tools effectively into EFL contexts. This study asked: To what extent do chatbot interactions promote sustained growth in linguistic complexity – operationalized through changes in lexical diversity measured by token type ratio, (TTR), mean sentence length, expanded responses, self-correction, follow-up questions, and total turns – among Japanese EFL learners?

Methods

Participants

Eighteen first-year English majors at a small private university in northern Japan took part in this study. All participants were enrolled in required English communication courses and gave informed consent for their interaction data to be analyzed. Although the sample size is modest, the study was designed as a longitudinal classroom investigation. The statistical analyses are therefore intended to identify developmental patterns within this instructional context rather than to support broad generalizable claims.

Design and procedure

The study followed a longitudinal design across two academic terms (Spring and Fall 2024). Each week, students completed a structured chatbot role-play task aligned with the units of *American Headway 1* and *American Headway 2* (3rd ed.; Oxford University Press). For each session, students were given a prompt introducing a new exchange-student persona corresponding to the unit topic. While the chatbot was consistently framed as an international exchange student, its academic background varied by chapter (e.g., an environmental studies major for a unit on environmental issues).

Students used ChatGPT via the smartphone app's voice interface and were instructed to engage in approximately 10 minutes of spoken dialogue. Prompts required students to initiate conversation, respond to follow-up questions, and incorporate target themes from the unit. Students submitted transcripts automatically generated by the app at the end of each session. No direct corrective feedback was provided on the chatbot exchanges during the term. This design allowed the chatbot interaction to function as independent communicative practice rather than as a teacher-mediated activity.

Data collection

The text of Chatbot interactions formed the primary dataset. End-of-term surveys were also administered to collect self-reported perceptions of learning, though these were not the focus of the present paper. Only student turns were analyzed; chatbot output was excluded from all measures.

Measures

From the logs, six metrics were calculated:

- ▶ **Lexical diversity** (Type-Token Ratio, TTR) per session. TTR was calculated as the ratio of unique word types to total tokens; because TTR is sensitive to text length, interpretation was made cautiously and considered alongside total turns and sentence length.
- ▶ **Mean sentence length** (average tokens per sentence).
- ▶ **Expanded answers**, defined as turns ≥ 5 tokens. (e.g., "I usually go to the gym after school because I want to stay healthy").
- ▶ **Self-correction count**, instances of revision or explicit correction within a turn (e.g., "I go to – sorry – I went to Kyoto last year").
- ▶ **Follow-up questions**, student turns ending with a question that extended the interaction (e.g., "What do you usually do on weekends?").
- ▶ **Total turns** per session.

Analysis

Changes were assessed using paired-sample *t*-tests comparing early and late sessions, as well as repeated-measures ANOVAs to identify trends across sessions. Analyses were run separately for spring and fall cohorts and then compared across semesters. Statistical significance was set at $p < .05$. Given the classroom-based design and

modest sample size, these analyses were intended to identify developmental patterns within this instructional context rather than to establish causal effects.

Results

Table 1 summarizes the statistically notable results for the spring and fall cohorts. As indicated in the table note, most measures did not demonstrate statistically significant change across sessions.

Spring 2024

Paired *t*-tests revealed a significant increase in total turns between the first and last sessions, $t(17) = 2.42$, $p = .028$, $d^z = 0.57$. Self-correction showed a marginal increase, $t(17) = 1.91$, $p = .074$, $d^z = 0.45$. Session-level ANOVAs indicated a significant upward trend in self-correction counts across the term, $F(10, 170) = 3.34$, $p < .001$, $\eta p^2 = .16$, suggesting increased metacognitive monitoring. Lexical diversity also trended upward but did not reach significance, $F(10, 170) = 1.87$, $p = .053$. Sentence length and expanded answers showed descriptive increases, though neither change was statistically significant.

Fall 2024

Paired *t*-tests showed a significant increase in lexical diversity, $t(17) = 5.51$, $p < .001$, $d^z = 1.30$, and a marginal increase in follow-up questions, $t(17) = 2.17$, $p = .055$, $d^z = 0.51$.

ANOVAs found no significant session-to-session differences across any measure in the fall term (all $p > .05$; e.g., $TTR F(10, 170) = 1.45$, $p = .164$). Mean sentence length increased slightly, while expanded answers declined over time.

Spring/fall comparison

Paired-sample *t*-tests found no statistically significant differences between spring and fall data for any metric (all $p > .05$). While no clear differences between semesters were detected, the results indicate that patterns of chatbot-supported language use were comparable across terms.

Summary of findings

Taken together, the results suggest that chatbot use was associated with early increases in interaction and self-correction in spring, while the fall data demonstrated some lexical diversity gains. However, vocabulary-related measures did not demonstrate consistent session-to-session growth, and expanded answers did not reliably increase over time. These patterns highlight both the potential and the limitations of repeated chatbot tasks in EFL contexts.



Table 1. Statistically notable effects across spring and fall terms

Metric	Test	Statistic (df)	p	Effect size
Total turns (Spring)	<i>t</i>	2.42 (17)	.028 *	<i>dz</i> = 0.57
Self-correction (Spring)	<i>F</i>	3.34 (10, 170)	< .001 ***	$\eta_p^2 = .16$
Lexical diversity (Fall)	<i>t</i>	5.51 (17)	< .001 ***	<i>dz</i> = 1.30
Follow-up questions (Fall)	<i>t</i>	2.17 (17)	.055 †	<i>dz</i> = 0.51

Note. $p < .05$ *, $p < .01$ **, $p < .001$ ***, † = marginal ($p < .10$).
All other metrics were nonsignificant. Repeated-measures ANOVAs: $k = 11$ sessions, $n = 18$; t tests paired, $df = 17$.

Discussion

This study provides longitudinal evidence of how structured chatbot interactions influence EFL learners’ language use over two academic terms. The results show both the promise and the limitations of chatbot-assisted practice. In spring, students significantly increased their number of turns and showed growth in self-correction, suggesting that chatbots can encourage more active participation and greater meta-cognitive awareness. In Fall, lexical diversity improved significantly, suggesting that continued chatbot use can promote vocabulary variety under certain conditions. However, expanded answers declined in the fall cohort, and overall gains tended to plateau, perhaps due to a lack of variation in task design.

These findings align with previous research suggesting that chatbots can facilitate initial linguistic improvements, particularly in interactional fluency and sentence elaboration (Ji et al., 2022; Zou et al., 2023). At the same time, the limited gains in lexical development mirror concerns raised in recent reviews that vocabulary growth is often constrained when tasks are repetitive or lack targeted support (Antony & Ramnath, 2023; Polakova & Klimova, 2024). The mixed results across semesters underscore the complexity of integrating chatbots into language education: while some dimensions of language use improve, others remain static or even regress when novelty fades.

Pedagogically, these findings suggest that chatbots may be particularly effective for initial engagement and structured communicative practice. In the present design, weekly 10-minute role-play tasks were aligned with textbook units and varied primarily through changes in persona and topic. While this provided thematic continuity, the conversational structure itself remained relatively consistent across sessions. To sustain development beyond initial gains, instructors may need to diversify the interactional demands of chatbot tasks, for example, by introducing problem-solving scenarios, debate formats, role reversals, or explicit lexical targets. Rotating conversational roles, increasing cognitive challenge, or requiring cumulative use of previously introduced vocabulary may help prevent stabilization of performance. Teacher mediation also remains important, not necessarily through direct correction, but through post-task reflection, goal setting, or strategic feedback that situates chatbot interaction within broader communicative objectives (Ballıdağ & Aydin, 2025). A blended approach that integrates chatbot practice with peer or teacher-guided discussion may help address the leveling-off patterns observed in the later sessions.

Several limitations must be noted. The study involved a relatively small sample ($n = 18$) at a single institution, which limits generalizability. The measures relied on surface-level metrics such as TTR and sentence length, which may not capture deeper syntactic or discourse-level development. In addition, count-based measures such as self-correction and follow-up questions may not meet normality assumptions; future analyses could model these as rates or with Poisson or negative-binomial regression to ensure robustness. Because TTR is sensitive to text length and total turns varied across sessions, future work should use length-robust lexical diversity measures (e.g., MTLT, HDD) or include total tokens as a covariate. Additionally, although the design spanned two terms, longer longitudinal studies are needed to examine retention and delayed effects. Future research should explore adaptive chatbot systems that have memory to better match learner progress and combine chatbot practice with human interaction to maximize authenticity. Larger-scale and multi-site studies would help clarify the conditions under which chatbots most effectively support sustained EFL development. In this context, the statistical analyses reported here should be understood as supporting reflective insight into classroom-based developmental patterns rather than confirmatory hypothesis testing.

Conclusion

This study examined how Japanese university students engaged with AI chatbots across two academic terms, focusing on objective measures of linguistic development. The results suggested potential benefits in interaction and self-correction in Spring, with trends – though not consistently significant – toward greater sentence elaboration, and gains in lexical diversity in the Fall term. At the same time, vocabulary-related measures did not demonstrate consistent session-to-session growth, and expanded answers declined in the Fall cohort, with several metrics stabilizing over time. These mixed findings indicate that chatbots may function most effectively as a supplement rather than a replacement for guided instruction. Overall, the results highlight the value of structured chatbot practice as an initial scaffold for EFL development while underscoring the importance of continued pedagogical mediation to sustain progress.

Use of generative AI

This manuscript made use of ChatGPT to assist in paraphrasing and improving the clarity of text as well as to direct the researcher in using Python for data analysis (see Table 1). No AI tools were used for the generation of research hypotheses or the writing of conclusions. The final manuscript was authored and reviewed by the listed author without additional AI input.

References

- Antony, S., & Ramnath, R. (2023). A phenomenological exploration of students' perceptions of AI chatbots in higher education. *IAFOR Journal of Education: Technology in Education*, 11(2).
<https://iafor.org/journal/iafor-journal-of-education/volume-11-issue-2/article-1/>
- Ballıdağ, M., & Aydin, S. (2025). A comparison of the effects of AI-based chatbots and peer interactions on speaking anxiety among EFL learners. *Future in Educational Research*, 3(2), 224–238. <https://doi.org/10.1002/fer3.70005>
- Chan, C., & Hu, W. (2023). Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education*, 20(1), Article 43.
<https://doi.org/10.1186/s41239-023-00411-8>
- Ji, H., Han, I., & Ko, Y. (2022). A systematic review of conversational AI in language education: Focusing on the collaboration with human teachers. *Journal of Research on Technology in Education*. Advance online publication.
<https://doi.org/10.1080/15391523.2022.2142873>
- Labadze, L., Grigolia, M., & Machaidze, L. (2023). Role of AI chatbots in education: Systematic literature review. *International Journal of Educational Technology in Higher Education*, 20(1), Article 56. <https://doi.org/10.1186/s41239-023-00426-1>
- MacWhinnie, S. G. B. (2025). Exploring the efficacy of ChatGPT's audio features in language learning. *JALTCALL Trends*, 1(1). <https://doi.org/10.29140/jct.v1n1.2145>
- Polakova, P., & Klimova, B. (2024). Implementation of AI-driven technology into education: A pilot study on the use of chatbots in foreign language learning. *Cogent Education*, 11(1), Article 2355385.
<https://doi.org/10.1080/2331186X.2024.2355385>
- Viktorivna, L. K., Oleksandrovych, A. V., Oleksandrivna, I. K., & Oleksandrivna, N. K. (2022). Artificial intelligence in language learning: What are we afraid of? *Arab World English Journal*, 8, 262–273. <https://doi.org/10.24093/awej/call8.18>
- Zou, B., Reinders, H., Thomas, M., & Barr, D. F. (2023). Editorial: Using artificial intelligence technology for language learning. *Frontiers in Psychology*, 14, Article 1287667. <https://doi.org/10.3389/fpsyg.2023.1287667>

Author biodata

Steven G. B. MacWhinnie has been living and working in Japan since 2007. He holds a Master's in TESOL from Shenandoah University. He is an associate professor in the English Department at Hirosaki Gakuin University. His research interests include AI, Timed Writing, and Motivation.