Koh, E., et al. (Eds.) (2026). Proceedings of the 2nd International Conference on Learning Evidence and Analytics. Asia-Pacific Society for Computers in Education.

Penny for Your Thoughts: EFL Learner Perceptions of GenAI Writing Support

Steve WOOLLASTON*, Brendan FLANAGAN, Yuko TOYOKAWA & Hiroaki OGATA
Graduate School of Informatics, Kyoto University, Japan
*s.m.woollaston@gmail.com

Abstract: This study investigates the perceptions of junior high school students in Japan using "Penny," a generative AI writing support chatbot. Utilising the extended Technology Acceptance Model (exTAM) and thematic analysis of qualitative feedback, the research evaluates learner perceptions and writing feedback preferences. Results indicate high levels of perceived ease of use and usefulness, with students valuing the immediacy of feedback and the friendly AI persona. However, a "proficiency mismatch" was identified, where the AI's linguistic output occasionally exceeded learner comprehension levels. Students had a strong preference for grammar and word choice feedback over global stylistic or organisational advice. These findings suggest that while GenAI is a useful tool for lowering communication anxiety and providing rapid corrective feedback, pedagogical success depends on matching AI output to learner proficiency levels.

Keywords: chatbot, EFL, L2 writing feedback, language learning, Technology Acceptance Model (TAM), Automated Writing Evaluation (AWE), LLM, GenAI

1. Introduction

The rapid advancement of Large Language Models (LLMs) has transformed the landscape of Computer-Assisted Language Learning (CALL), shifting the role of AI from simple rule-based error checking to sophisticated, conversational tutors. For English as a Foreign Language (EFL) learners in Japan, who often face high levels of communication anxiety and limited opportunities for authentic output, generative AI (GenAI) chatbots offer a low-stakes environment for immediate, personalised writing feedback. However, while the technical capabilities of these tools are well-documented (Slamet & Basthomi, 2025), their utility and integration into teaching and learning also depends largely on student preferences and subjective user experience. Understanding which features learners value and which they find obstructive is essential for developing AI tools that support learning and sustain engagement. This study investigates these research questions:

RQ1: How do learners perceive *Ease of Use*, *Usefulness*, and *Enjoyment* of the writing support chatbot? What are their *Attitudes* toward, and do they have *Intentions* to continue practising with the chatbot?
RQ2: What are learners' impressions of the strengths and weaknesses of the chatbot?
RQ3: Which types of feedback (e.g., grammar, word choice, spelling, organisation) do learners find most useful and least useful?

2. Related Work

Research on language learning chatbots utilising the Technology Acceptance Model (TAM) indicates that utility and ease of use are important determinants of students' intention to continue using AI tools (Chang et al., 2021). Learners generally view chatbots as user-friendly and accessible (Mohamed & Alian, 2023). However, chatbot usage over time often wanes, potentially due to the voluntary nature of usage or the "novelty effect", where student engagement drops after the initial excitement of the technology fades (Fryer et al., 2019). Furthermore, while some students find AI feedback engaging, others prefer human instruction for a variety of reasons, including lack of trust and the perceived inability of AI to provide nuanced, socio-emotional support (Escalante et al., 2023).

Chatbots offer distinct advantages over traditional classroom instruction, primarily through their availability as “tireless assistants” providing 24/7 language practice (Huang et al., 2022; Woollaston et al., 2024). They provide a low-anxiety environment where learners can practice without fear of peer or teacher judgement, which is particularly beneficial for reducing foreign language anxiety (Wiboolyasarin et al., 2025). Chatbots can give immediate, personalised feedback, allowing students to correct mistakes in real time (Lin & Crosthwaite, 2024). GenAI chatbots possess vast knowledge bases and the sophisticated capacity to process nuanced, context-dependent language. This provides students with authentic, open-ended conversational practice that mirrors real-world interaction more closely than previously (Mun, 2024). Despite their utility, chatbots face significant limitations. Proficiency barriers exist: lower-level learners may struggle to utilise the technology effectively or be unable to comprehend complex, lengthy AI responses, leading to cognitive overload or frustration (Duong & Chen, 2025; Rong et al., 2025). Chatbots have been criticised for lacking the "human touch," failing to understand cultural nuances or provide the deep, empathetic support found in teacher-student interactions (Lin & Crosthwaite, 2024). There are also concerns regarding accuracy, as chatbots can generate invalid or hallucinatory responses (Mustaffa et al., 2025).

Research shows that students consistently demonstrate a preference for direct and explicit feedback on grammar, syntax, spelling, and punctuation, viewing these surface level corrections as essential for linguistics accuracy and avoiding negative grade outcomes (Lin & Crosthwaite, 2024). They also value feedback on word choice, particularly when it helps them clarify meaning or expand lexical range (Bobrova, 2018). Preferences on organisation are mixed, although some students focus primarily on local errors, many - especially at higher proficiencies - actively seek guidance on how to structure their work to improve flow (Zou et al., 2025). Students generally favour feedback that balances constructive criticism with positive reinforcement (Underwood & Tregidgo, 2006).

3. Methodology

3.1 *Participants and System Design: The Penny Chatbot*

The study was conducted at a high-performing public junior high school in Japan, involving 115 EFL learners across three classes. The participants, aged 14 to 16, ranged in English proficiency from CEFR A1 - B1 level. The cohort had limited prior experience with generative AI tools in a formal classroom setting. The learners interacted with "Penny," a custom-built English writing support chatbot powered by the OpenAI `gpt-4o` API. Penny was wrapped in a dedicated educational persona via a system prompt designed to act as a supportive tutor rather than a simple error corrector. Each day, a new writing prompt would become available for student response. Students could also review their previous work and feedback at any time. The system included the following features:

- **Language:** Interface and feedback explanations were provided in English or Japanese; student writing submissions were restricted to English only.
- **Persona:** The LLM was instructed to be "friendly, patient, and supportive," with responses limited to approximately 50 words to prevent cognitive overload.
- **Gated feedback:** A minimum of 100 characters were required before the *Check my writing* button became active, ensuring a sufficient writing sample for feedback.

3.2 *Procedure*

The study spanned four months. Students used the chatbot during regular English lessons approximately three to four times per week. Each session followed a consistent routine:

1. **Peer discussion:** A five-minute *Think-Pair-Share* activity where students verbally brainstormed ideas for the daily prompt (e.g., "How do you help your family at home?").
2. **Writing and feedback:** A 10-20 minute independent writing period where students drafted text, received feedback, and revised their work. The teacher was able to see students’ writing through a dashboard, support struggling students, showcase specific

student work, or provide class instruction on common errors in the students' writing.

3.3 Data Collection and Analysis

Data was collected via an anonymous Google Form administered in late March. The survey was given in Japanese and English to ensure understanding:

- **exTAM**: Items were adapted from the exTAM on a 7-point Likert scale (1 = Strongly Disagree, 7 = Strongly Agree). These items measured the domains: *Perceived Ease of Use, Usefulness, Enjoyment, Attitude*, and *Intention to Use* (Wu & Gao, 2011).
- **Open questions:** Open-ended questions asked learners to justify their choices, identify the chatbot's strengths, and describe specific improvements they wanted in the system.
- **Feedback preferences**: Students identified the "most useful" and "least useful" types of feedback from six categories (*Tone and Style*, *Word Choice*, *Grammar and Syntax*, *Spelling and Punctuation*, *Structure and Organisation*, and *Clarity and Coherence*).

Qualitative data obtained from the open-ended survey questions regarding the chatbot's strengths and areas for improvement were collaboratively analysed and classified by two researchers (both experienced in EFL teaching and learning) using an inductive thematic analysis (Braun & Clarke, 2006) approach. The analysis proceeded through three phases:

1. **Open coding**: Initial labels were assigned to distinct semantic units within the responses (e.g., assigning "fast reply" and "immediate answer" to a preliminary code)
2. **Theme development**: Where related codes were synthesised into broader, overarching categories (e.g., *Immediacy of Feedback*)
3. **Quantification**: Themes summated to identify the dominant themes.

4. Results

4.1 *System Usage and Interaction Metrics*

During the semester, students engaged in 4,651 writing sessions and exchanged over 21,000 messages with the Penny chatbot. While the average learner completed 39 sessions and sent 177 messages, the high standard deviations (12.57 and 156.53, respectively) highlight a broad spectrum of engagement. Notably, 98% of student interactions were in Japanese rather than English.

Table 1. *Interaction Summary (n=119)*

Metric	Mean	SD	Median	Min	Max
Writing sessions per learner	39.06	12.57	40	4	101
Total messages per learner	176.92	156.53	137	7	917
Messages per session	5.58	6.96	4	1	116
Character count (initial draft)	210.05	151.54	177	100	3000
Character count (final submission)	183.66	166.54	167	0	3000

Initial drafts averaged 210 characters, but decreased to an average of 183.66 characters at the end of a session. This contraction suggests that Penny's feedback encouraged editing rather than expansion. Log data also revealed a skewed temporal distribution: while the average session lasted 32 minutes, the median was only seven minutes. This indicates that most interactions were brief check-ins, though a subset of students engaged in much longer sessions. Overall, student writing was assessed at the A2–B1 level.

4.2 *exTAM (RQ1)*

Learners reported generally positive perceptions of Penny, with high mean scores across all exTAM subscales (Table 2). PEOU, PU, and ATT received the highest ratings (means > 5.4), indicating high perceived accessibility and utility; PE and INT were lower but remained above the neutral midpoint, with high internal consistency across all measures ($\alpha > 0.87$).

Table 2. *exTAM subscale ratings (n=115)*

	Mean	SD	Minimum	Maximum	α
PEOU	5.57	1.09	1	7	0.88
PU	5.46	1.21	1	7	0.94
PE	4.99	1.38	1	7	0.98
ATT	5.45	1.22	1	7	0.88
INT	4.70	1.20	1	7	0.90

4.3 *Chatbot Strengths and Weaknesses (RQ2)*

4.3.1 Perceived Strengths

Analysis of the open-ended responses regarding the system's strengths revealed that learners (*n*=88) primarily valued its functional utility, with *Correction and Accuracy* (26.1%) and *Learning and Skill Acquisition* (19.3%) emerging as dominant themes. Students frequently commended the system's precision, noting that it "wipes out mistakes" and "corrects sentences that I thought were okay." The *Immediacy of Feedback* (13.6%) was highlighted as a key advantage over traditional methods; one student explained, "It can check fast, so I can review right now," while another appreciated that "the correct answer comes immediately." Beyond simple correction, learners recognised the tool's role in scaffolding their language development, stating that it "teaches more natural expressions" and helped them "learn many words" and "get used to writing long sentences."

The user experience and affective dimensions were also highly rated, with *Persona and Affect* (13.6%), *Clarity of Explanations* (11.4%), and *Usability* (11.4%) forming a significant portion of the positive feedback. Students frequently anthropomorphised the system, describing it as "kind," "friendly," and "patient." This affective dimension appeared to lower the threshold for asking questions. One student explicitly compared the interaction to a peer relationship: "It chats with me about daily things and feels like a friend." Another noted, "Penny was kind when answering questions." The visual design also contributed to this perception, with comments such as "the icon is very cute" and "unique." This approachability was supported by the system's bilingual capabilities, which students found "easy to understand" because it "teaches specifically one by one." The ease of access further encouraged casual engagement, with one respondent valuing the ability to "ask directly anytime about what I don't understand."

4.3.2 Desired Improvements

Regarding potential improvements (*n*=87), the analysis revealed that *Pedagogical Feedback Quality* was the most frequent concern, accounting for 47.1% of all comments. Within this category, learners frequently cited the difficulty of the vocabulary used by the chatbot as a barrier, with one student noting, "The words are difficult, so it would be interesting to have levels like '2nd year student'." Another significant pedagogical issue was feedback loops, where the system provided contradictory or circular advice; as one participant described, "I rewrite it as Penny said, but then it changes it... I don't know which is better." Additionally, concerns regarding accuracy were raised, where learners felt the system "corrects too much, and sometimes the meaning of my sentence changes," or flagged grammatically correct sentences as erroneous.

The second most prominent area for improvement related to *System Constraints*. Eleven students (11.5%) expressed frustration with the minimum character requirement, stating, "I want to be able to use it with short sentences" or noting that the requirement was difficult to meet within the limited class time. *Technical Performance* issues, such as latency and connection errors, were mentioned by eight learners (9.2%); "Sometimes, the answer is slow." Finally, 16.3% of comments focused on *Feature Requests*, specifically asking for translation support, "I want to translate to Japanese when I don't understand a word", and

more customisable persona settings, with suggestions to "make it a bit cuter" or include "family structure or age" to deepen the roleplay.

4.4 *Feedback Utility (RQ3)*

Participants demonstrated a distinct preference for form-focused corrections over global or stylistic advice (Table 3). *Grammar and Syntax* (33.91%) and *Word choice* (25.22%) were identified as the most beneficial categories, cumulatively accounting for nearly 60% of the responses. Conversely, *Tone and Style* was the most frequently cited as the least useful category (26.96%), indicating that abstract feedback was a significantly lower priority for these learners than immediate mechanical and lexical accuracy.

Table 3. *Feedback Type Preference (n=115)*

Feedback Type	Most Useful		Least Useful	
	n	%	n	%
Grammar and Syntax	39	33.91	7	6.09
Word choice and Vocabulary	29	25.22	14	12.17
Spelling and Punctuation	18	15.65	17	14.78
Structure and Organisation	14	12.17	15	13.04
Clarity and Coherence	10	8.70	15	13.04
Tone and Style	4	3.48	31	26.96
Other	1	0.87	16	13.91

5. Discussion

The findings of this study suggest that generative AI chatbots, when wrapped in a supportive pedagogical persona, are viewed as highly effective and accessible tools for Japanese EFL learners. Consistent with previous research applying the TAM to CALL (Chang et al., 2021), participants rated PEOU and PU highest among all metrics. This indicates that the "Penny" system successfully lowered technical barriers and friction, allowing students to focus on writing rather than interface navigation. The qualitative data reinforces this, as students frequently praised the "immediacy" of feedback, a clear advantage for students and busy teachers. The high Attitude scores, coupled with comments describing Penny as "kind" and "patient," point toward the potential for anthropomorphic design to reduce the anxiety often associated with teacher or peer evaluation (Wiboolyasarin et al., 2024).

However, a discrepancy emerged between the high perceived utility and the slightly lower Intention to Use. While students recognised the chatbot's value for error correction, the "shrinking" of drafts (from 210 to 183 characters) suggests that learners may view the tool primarily as a proofreader rather than a collaborative writing partner. This aligns with the feedback preferences observed: learners overwhelmingly favoured explicit corrections on *Grammar and Syntax* (33.9%) and *Word choice* (25.2%) over abstract categories like *Tone and Style* (3.4%). This preference for surface-level accuracy is typical of A1–B1 learners who prioritise linguistic correctness over stylistic nuance (Lin & Crosthwaite, 2024).

Despite the positive reception, pedagogical friction points remain. Nearly half of the qualitative suggestions for improvement focused on feedback quality, specifically the difficulty of Penny's suggested vocabulary and grammar, and circular feedback loops. This highlights a critical "proficiency mismatch": while the persona was friendly, the LLM's linguistic output occasionally exceeded the learners' comprehension levels, leading to cognitive overload and frustration, rather than scaffolding.

This study is limited by its reliance on self-reported perceptions and system logs without a concurrent measure of longitudinal linguistic acquisition; thus, we cannot confirm if high engagement translates to durable language proficiency gains. Additionally, the study was conducted at a single high-performing junior high school, which may limit generalisability to other proficiency contexts. Future research should investigate adaptive prompting strategies that restrict AI feedback to specific CEFR levels (e.g., A2 - B1) to

resolve the difficulty mismatch. Furthermore, longitudinal studies are required to determine if the *Intention to Use* stabilises once the "novelty effect" of the technology subsides.

Acknowledgements

Many thanks to the teacher and students who participated in this research, which was supported by Council for Science, 3rd SIP JPJ012347, and JSPS Grant-in-Aid for Scientific Research (B) JP20H01722 and JP23H01001, (Exploratory) JP21K19824, JP22KJ1914, (A) JP23H00505, and NEDO JPNP20006.

References

Bobrova, L. (2018). The effects of written feedback on ESL writers' ability to edit word choice errors. *International Journal of Applied Linguistics & English Literature*, *7*(3), 1.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101.

Chang, T.-S., Li, Y., Huang, H.-W., & Whitfield, B. (2021). Exploring EFL students' writing performance and their acceptance of AI-based automated writing feedback. *2021 2nd International Conference on Education Development and Studies*. Hilo HI USA.

Duong, T.-N.-A., & Chen, H.-L. (2025). An AI chatbot for EFL writing: Students' usage tendencies, writing performance, and perceptions. *Journal of Educational Computing Research*.

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*(1), 57.

Fryer, L. K., Nakao, K., & Thompson, A. (2019). Chatbot learning partners: Connecting learning experiences, interest and competence. *Computers in Human Behavior*, *93*, 279–289.

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.

Lin, S., & Crosthwaite, P. (2024). The grass is not always greener: Teacher vs. GPT-assisted written corrective feedback. *System*, *127*, 103529.

Mohamed, S. S., & Alian, E. M. (2023). Students' attitudes toward using chatbot in EFL learning. *Arab World English Journal For Translation and Literary Studies*, *14*(3), 15–27.

Mun, C. Y. (2024). EFL learners' English writing feedback and their perception of using ChatGPT. *Journal of English Teaching through Movies and Media*, *25*(2), 26–39.

Mustaffa, N. E., Lai, K. E., Preece, C. N., & Wong, F. Y. (2025). A bibliometric review of large language model hallucination. *International Journal of Research and Innovation in Social Science*, *9*(9), 5025–5037.

Rong, M., Yao, Y., Li, Q., & Chen, X. (2025). Exploring student engagement with artificial intelligence-guided chatbot feedback in EFL writing: interactions and revisions. *Computer Assisted Language Learning*, 1–30.

Slamet, J., & Basthomi, Y. (2025). Examining the challenges and opportunities of ChatGPT in EFL education: A systematic literature review. *Journal of University Teaching & Learning Practice*, *22*(2). https://doi.org/10.53761/deezkh88

Underwood, J. S., & Tregidgo, A. P. (2006). Improving student writing through effective feedback: Best practices and recommendations. *The Journal of Teaching Writing*, *22*, 73–98.

Wiboolyasarin, W., Wiboolyasarin, K., Tiranant, P., Boonyakitanont, P., & Jinowat, N. (2024). Designing chatbots in language classrooms: an empirical investigation from user learning experience. *Smart Learning Environments*, *11*(1).

Wiboolyasarin, W., Wiboolyasarin, K., Tiranant, P., Jinowat, N., & Boonyakitanont, P. (2025). AI-driven chatbots in second language education: A systematic review of their efficacy and pedagogical implications. *Ampersand (Oxford, UK)*, *14*(100224), 100224.

Woollaston, S., Flanagan, B., & Ogata, H. (2024). Chatbots and EFL learning: A systematic review. *Joint Proceedings of LAK 2024 Workshops*, 89–98.

Wu, X., & Gao, Y. (2011). Applying the extended technology acceptance model to the use of clickers in student learning: Some evidence from macroeconomics classes. *American Journal of Business Education*, *4*, 43–50.

Zou, S., Guo, K., Wang, J., & Liu, Y. (2025). Investigating students' uptake of teacher- and ChatGPT-generated feedback in EFL writing: a comparison study. *Computer Assisted Language Learning*, 1–30.