Exploring the Use of Generative AI Chatbots for Second Language Writing Feedback with Japanese University Learners of English as a Foreign Language

Sean A. WHITE

▶Key words

generative artificial intelligence (AI);
chatbot; ChatGPT;
second language writing;
English as a foreign language

▼Abstract

The emergence of generative AI (genAI) chatbots like ChatGPT presents both opportunities and challenges for second language teaching and learning. This is especially the case for writing due to AI's ability to quickly produce natural human-like texts by means of simple prompts. The present study explored the use of genAI chatbots not as composing but as feedback tools used by students on texts they wrote for an oral presentation. Participating students (N = 6) answered post-task questionnaires and submitted their draft and revised texts as well as chat logs for analysis. Results indicated that genAI could be used cautiously as a source of reference in addition to instructor feedback. In light of use- and feedback output-related issues raised, the need to develop AI literacy and better understanding of genAI through further research are highlighted.

1. Introduction

In November 2022, OpenAI released its "user-friendly," artificial intelligence chatbot application ChatGPT (OpenAI, 2022). It was widely recognized for its natural language processing (NLP) abilities using simple conversational prompts rather than computer code.

These abilities were based on the GPT (generative pre-trained transformer) large language model (LLM) GPT-3, an artificial neural network built with human feedback-reinforced machine learning using a large and diverse dataset of human language (first available via API in 2020; OpenAI, 2020). While computer text generation of individual sentences based on predictive algorithms had already existed for some time (familiar to most people now through text-completion suggestions of mobile devices and email and word-processing programs), the transformer model incorporated by OpenAI into its GPT family of models allowed for greater "attention" and context awareness to not only construct individual grammatically correct sentences but entire natural-sounding, human-like texts with a high degree of rhetorical consistency (Godwin-Jones, 2022, 2023). In effect, it appeared to be able to not only produce texts but also, because of its dialogic interface, communicate with users based on the messages (prompts) they entered. The fact that this ability was based on large amounts of information on the internet in its training data (including data from CommonCrawl, book corpora, and Wikipedia; Brown et al., 2020; Zhao et al., 2023) meant that ChatGPT was able to provide clear (if not always accurate) answers to user questions typically found through more involved internet searches, thus giving it an air of authority and perceived useability/productivity of a "super tool" with potential revolutionary uses.

However, ChatGPT and similar generative artificial intelligence (genAI) chatbots are fundamentally language-processing rather than information-processing tools (Walters & Wilder, 2023). As a result, the limitations of the technology became apparent: information was not up-to-date and limited to the cutoff date of the training data, and even then, due to its "emergent" nature it could "make things up" or "hallucinate" information and ideas in achieving its intended purpose of completing natural-sounding human language texts (OpenAI, 2022). Initially having a training cutoff date of September 2021 for the first-released and initial free version, GPT-3.5 (OpenAI, n.d.-b), the paid version, ChatGPT Plus, provided access to an updated model, GPT-4, with a more recent December 2023 training cutoff date (OpenAI, n.d.-b). GPT-4 also showed increased information accuracy and reduced instances of hallucinations (OpenAI, n.d.-a) and remains the most advanced model reserved for ChatGPT Plus users while free and paid users both have access to the better-optimized GPT-4o, released May 2024 (OpenAI, 2024). GPT-4 and GPT-4o have a number of other features, including multimodal (not just text) input, including file uploads, and can, via modifications first through plugins and now through custom instruction "GPT"s, connect to the internet for more recent information and services. The pace and degree of technological change mean that we will likely see continued and rapid improvement (Godwin-Jones, 2022).

The emergence of ChatGPT and similar LLM-based chatbots such as Anthropic's Claude, Google's Gemini (formerly Bard), and services that use OpenAI's GPT models such as Microsoft's Copilot (formerly Bing Chat) have generated a large degree of excitement for their potential applications—both positive and negative. ChatGPT's ability to generate not only text, but also images, computer code, and results of data analysis, all multimodally, have

had many predicting increased productivity and innovation. On the negative side, there is concern about job loss, deskilling, malignant uses to deceive others for financial, political or other gains, and accidents and other kinds of misfortune brought on by ceding too much of our lives to machines. In education, generative artificial intelligence chatbots have been heralded as "game-changers" for their potential to provide customized adaptive learning, and targeted for banning as tools that promote plagiarism and decreased thinking and intellectual accountability (Cotton et al., 2023; Fütterer et al., 2023; Kasneci et al., 2023; Lampropoulos et al., 2023; Moorhouse et al., 2023; Perkins, 2023; Sullivan et al., 2023).

As language processing tools based on LLMs, genAI chatbots and related technology are well-suited to being applied to second language education in particular. Researchers have identified a number of affordances and pitfalls, including potential benefits of individualized and adaptive learning, learner autonomy and efficiency, increased engagement and motivation, and language and educational equity, but also problems related to plagiarism and academic integrity, authenticity, training and use biases, inequality, and premature- and over-reliance leading to inhibited skill development (Barrot, 2023a; Pack & Maloney, 2023; Warschauer et al., 2023). The sudden appearance of this technology on a wide scale, however, means best pedagogical practices are yet unclear, though research interest has, unsurprisingly, surged since November 2022 (a Google Scholar search for OpenAI shows nearly a tripling of results returned from 2022 to 2023, a number that was already surpassed by August 2024). Based on this as well as a lot of trial and error, there appears to be some initial consensus. First, wholescale outsourcing of learning and especially text creation to genAI is unacceptable and, unsurprisingly, many institutions quickly banned or limited their use for academic work (Fütterer et al., 2023; Moorhouse et al., 2023; Sullivan et al., 2023). However, the use of AI to support learning has been broadly recognized. This includes as a language-resource, practice, thinking and brainstorming, and feedback tool used by both learners and teachers (Barrot, 2023a, 2023b; Pack & Maloney, 2023; Warschauer et al., 2023). At present, details regarding how and to what extent these are best implemented in terms of pedagogy and other issues are unclear and there is much need for research following the initial studies that have begun to appear.

The present study is an exploratory investigation into the use of genAI chatbots, including ChatGPT, in an oral skills development class in an English as a foreign language/lingua franca, communication-focused course of study for social science majors. It consisted of the use of genAI by students in preparing for a composed oral presentation modeled in part on TED Talks that served as learning material over the course of a semester.

2. Background & Research Questions

Rationale for inclusion in second language learning and writing

A number of articles following the release of ChatGPT have argued for its use in second language learning. Mizumoto (2023) recognized the relative accessibility/user-friendliness of genAI such as ChatGPT in comparison to corpora which it can complement. He positioned it as one of many data-driven learning (DDL) resources such as concordance tools, machine translators (MT) (e.g., DeepL and Google Translate), corrective feedback tools (e.g., Grammarly), and other text generators (e.g., Quillbot) in a proposed Metacognitive Resource Use framework and pedagogical approach which seeks to develop learner metacognitive knowledge and regulation via direct instruction. Godwin-Jones (2022) similarly emphasized the need for knowledgeable instructors, making it incumbent on language educators to work to clarify appropriate uses of AI tools and creatively incorporate them into their instruction whenever possible.

Warschauer et al. (2023) argued forcefully for the affordances of AI text generators as empowering tools for second language writers (and learners) and the need to incorporate them into instruction. This requires experimentation by instructors to develop best practices while mitigating the real threats of inauthenticity, plagiarism, and insufficient language skill development. The authors proposed a framework for second language writer AI literacy based in part on Ng et al. (2021) that includes understanding, using ("accessing" and "prompting"), corroborating, and incorporating text generators and their output. Following this, for using ChatGPT, they listed specific writing functionalities of content development (outlining, brainstorming), corpus search (example sentences, vocabulary, grammar), text modification (adjustment of text difficulty level, genre, tone, etc.), feedback and revision (proofreading, editing, analysis and explanation of errors), response to language-related questions, support for sentence generation (translation, templates and sentence starters), vocabulary/grammar support, and paraphrasing and summarizing (p. 12).

Su et al. (2023) argued that ChatGPT is well suited for learning argumentative writing due to its dialogic nature and ability to help create and assess outlines, provide example sentences and texts, answer questions, provide feedback on learner-submitted texts, and proofread and edit language. The authors provided suggestions and examples of uses throughout the writing process, including outline preparation, content revision, proofreading, and post-writing reflection. In particular, they highlighted the potential for timely feedback on outlines and drafts in terms of content and proofreading. In their testing using GPT-3.5, they found content feedback to be largely evaluative rather than elaborative, with some inconsistency with multiple submissions. However, grammar checking of individual sentences was accurate and the chatbot was able to provide alternatives based on its interpretation of

written content. Full-text checking resulted in grammatical, mechanical, lexical, and full-sentence corrections that the authors determined produced a better text overall. In addition to both Su et al. and Godwin-Jones, and echoing Mizumoto and Warschauer et al., Kong et al. (2024) presented a reflective framework for incorporating ChatGPT as a part of process writing.

Research on ChatGPT assessment and feedback capabilities

In addition to the above research that provides a rationale for including genAI in second language learning and writing, particularly for feedback on learner-composed texts, there have been several studies to evaluate ChatGPT's abilities to assess texts (Baffour et al., 2024; Liang et al., 2023; Mizumoto & Eguchi, 2023; Pfau et al., 2023). Generally, it was found in these studies to be accurate to a certain extent for overall ratings that correlated well with human ratings, and which improved with more specific language criteria (Mizumoto & Eguchi), but became much less valid and reliable than human assessment when assessing individual components such as grammar, organization, etc. (Baffour et al.; Pfau et al.). As a feedback tool, Liang et al. found 30% overlap with feedback provided by human reviewers of research papers submitted to journals, which was consistent with usual overlap levels between human reviewers. Lastly, Wu et al. (2023) examined grammar correction abilities of ChatGPT in comparison to Grammarly and a specialized corpus-based tool, GECToR, and found it able to catch more errors and have lower false negative results than either of the other two tools but at the expense of a larger number of false positives (flagged items that were not in fact errors), with measures of accuracy (precision) and recall decreasing with longer texts. The specialized tool had the fewest number of false positives, giving it higher precision at the expense of a larger number of false negatives. Although Grammarly was rated highest based on a balance of precision and recall, the authors concluded that there was promise for the use of ChatGPT for grammar error correction of texts, particularly when used in conjunction with other tools, which is consistent with Mizumoto (2023).

ChatGPT applied research

A small number of early experimental, quasi-experimental, and mixed-methods studies informed the current study. Escalante et al. (2023) conducted a study involving 48 English language learners of B1 and above proficiency where ChatGPT-4 was used by instructors to provide feedback on content and language use on weekly writing assignments. No difference was found between treatment group and control group receiving instructor feedback as measured on pre- and post-tests, with both groups exhibiting positive gains. Additionally, questionnaires showed participants as evenly divided in terms of preference for AI and

instructor feedback. Students preferring AI feedback emphasized its detail, accuracy, and clarity as well as convenience. Some students also indicated a preference for a combination of AI and instructor feedback, noting the ability to interact with tutors to discuss writing while also having access to the detailed AI feedback. The ability of AI to save instructors' time in providing their own feedback was noted as a benefit of using AI. Given the lack of group difference between learning outcomes as well as the preference division, the authors concluded that ChatGPT can be used for feedback without any loss in learning outcomes in comparison to instructor feedback but that it should be used in combination with the latter.

Guo and Wang (2023) compared feedback on content, organization, and language obtained by ChatGPT-3.5 vs. feedback generated by instructors on 50 essays written by Chinese learners of English of B2–C1 proficiency. They found both quantitative and qualitative differences between the two. First, ChatGPT produced more than two times the amount of feedback than instructors, distributed evenly between content, organization, and language, compared to instructors' greater focus on content and language. Notably, 18% of the ChatGPT feedback was determined to be "off-task." Second, ChatGPT produced significantly more directive feedback with specific suggestions as well as more praise, while instructors produced significantly more informative and query feedback prompting writers to consider action for themselves. In addition, ChatGPT offered more text summary compared to instructors, which the authors suggest may be beneficial to learners to consider organization of their texts. Instructor perceptions of ChatGPT feedback were generally positive though concerns were raised regarding the length and comprehensibility of feedback, especially for lower proficiency learners. The authors concluded that ChatGPT has the potential to complement instructor feedback and lessen burdens on teachers while indicating the need to appropriately address concerns and issues (through prompting and understanding of AI feedback) such as those just mentioned.

Lastly, Song and Song (2023) took a social constructivist position and conceived of ChatGPT as a potential "knowledgeable peer" able to support learners' writing and language development. In their experimental study involving 50 university students learning English in China, students in the treatment group used ChatGPT as a writing coach and tutor to provide explanations, examples, and feedback on their writing over a 12-week period. Writing authenticity and integrity were emphasized and plagiarism discouraged on weekly writing assignments. Students were also asked to reflect on their learning and use of ChatGPT in a writing portfolio. Results of the post-test and a questionnaire indicated significant improvements in writing skills and motivation compared to the control group.

Research questions

Given both the potential and issues related to the use of genAI in learning second language writing, the present study was conducted as an exploratory study with the following

research questions:

1. How can genAI be used as a feedback tool to support English language learners' process writing?
2. What prior experience do students have using genAI?
3. How do students use genAI as a feedback tool for checking their writing?
4. How do they perceive the process and results of using genAI as a feedback tool for checking content, organization, and language in their writing?
5. How accurate and useful are the writing feedback results provided by genAI and are they similar to or different from human/instructor feedback?
6. How do students use the results obtained from genAI as a feedback tool for checking their writing?

3. Method

Participants & procedure

The class in which the study took place consisted of nine (six male and three female) second-year students majoring in economics, business, law, and policy studies during the fall semester of the 2023 academic year. It was a 15-week oral skills class at the B1 level of English proficiency (as indicated by the results of the class placement test). The class focused on conversational, oral academic, and vocabulary and overall language proficiency development. Basic lessons involved the use of an oral skills textbook. At the completion of each unit in the textbook, students viewed TED Talks that were thematically connected to textbook unit topics. TED Talks were used as authentic examples for language input, listening practice, vocabulary reinforcement, as well as output practice based on talk transcripts and content reaction. They also served in part as model talks for more structured oral communication/presentations. Students watched a total of three talks over the semester. As an end-of-semester activity assignment, students were asked to create their own five-minute talk modeled on the TED Talks they had previously viewed in class. They were asked to submit possible topics for their talks on the university LMS. There was no restriction in terms of topic. Students were told they could (but were not required to) use genAI in brainstorming their topics and to submit their chat history in the case that they used any. Earlier in the semester, students were introduced to genAI and its possible uses for language study. Specifically, they were introduced to the free and paid subscription-based tool Poe (Quora, Inc., n.d.), an online and mobile and desktop application offering access to a number of chatbots, including ChatGPT, Claude, and Gemini as well as user-created bots. Students were given an opportunity to chat with chatbots to learn about English stative and non-stative verbs encountered in a lesson in the textbook by asking for explanations, example

sentences, and translations (these were compared to the textbook and other online sources for checking accuracy). After topics were submitted and confirmed, students were asked to submit the text for their talk on the LMS by the deadline indicated. They were instructed that talks should have some kind of clear message that the student wanted to share. In addition, they were explicitly told that the use of genAI or machine translation (MT) was not permitted to write their drafts.

Texts were written outside of class. After they were submitted, in class students were shown a section of a YouTube video (Vox, 2023) on the issues surrounding the use of AI in education which touched on AI literacy and possible helpful and harmful uses. Following the video and in line with its content, students were told that chatbots could be used as a helpful tool for feedback on their written texts. Taking advantage of the dialogic nature of chatbots, students could submit their texts and prompt and ask questions about what they had written in terms of clarity of message, organization, and flow of their texts as well as suggestions for improvement. An online resource was introduced with an initial prompt and questions to ask in both English and Japanese, with Japanese suggested as the potentially better option for understanding responses. The initial prompt was as follows (Figure 1):

Initial AI chat prompt (enter into chatbot):
I'd like to *have a chat about something I have just written. I will enter the text and then after you confirm, I will ask you a series of questions about it. You will act as a writing tutor and respond to each question I ask. OK?*
Enter text:
First, the text: " "
[Copy and paste your text + return. Do not include your name or other personal information.]
Questions (enter each, one by one):. . .

Figure 1 *Model Initial Chat Prompt for Content Feedback Included in Resource Provided to Students*

Table 1 summarizes the question-prompts provided, which were drawn in large part from writing rubrics found in Dzekoe (2017) and White (2023). Figures 2 and 3 show example AI content feedback responses to student-submitted texts.

Table 1 *List of All Text Content Feedback Prompt Questions Included in Resource Provided to Students*

No.	Prompt
1.	What is the main message of this text? Is it explicitly stated, or is it implied? Is there a clear topic sentence or conclusion that indicates the main message?
2.	Rhetorically, is the main message presented deductively, inductively, abductively, or some other way?
3.	What is the main purpose of the text? Is this explicitly stated, or is it implied?
4.	Is the main message sufficiently explained, described, and illustrated? What improvements can be made to make it more effective?
5.	What other minor ideas are expressed in the text? Do they contribute to or detract from the interest and usefulness of the text?
6.	What is the significance of the main message and other information and ideas in the text? Is it explicitly stated, or is it implied? How/why are the main message and other information and ideas useful, helpful, interesting, etc.? Does it describe its real-world relevance/applications?

7. What information or content is missing from the text? What else could or should be included?
8. Does the text appear credible? Are reliable sources referred to, either explicitly or implicitly?
9. Is the text organized appropriately and logically?
10. Is the text well-written in terms of grammar, vocabulary, register, mechanics, etc.?
11. Is the reasoning in the text clear and logical? Does the text exhibit critical thinking?
12. Is the text balanced? Does it exhibit any cognitive biases? Are counter-arguments, alternate perspectives and interpretations, or limitations acknowledged or discussed?
13. Does the text make any recommendations or indicate future directions? Are they explicit or implied? Are they appropriate, based on the text? Do they suggest any actions for others (listeners, readers) to take?
14. What questions might someone ask about this text? How might the author respond to these?
15. Is the text interesting and engaging? With whom would it resonate? Does it include stories, anecdotes, and/or interesting examples? Does it use rich language and/or rhetorical questions or other devices?
16. What other suggestions do you have to improve this text?

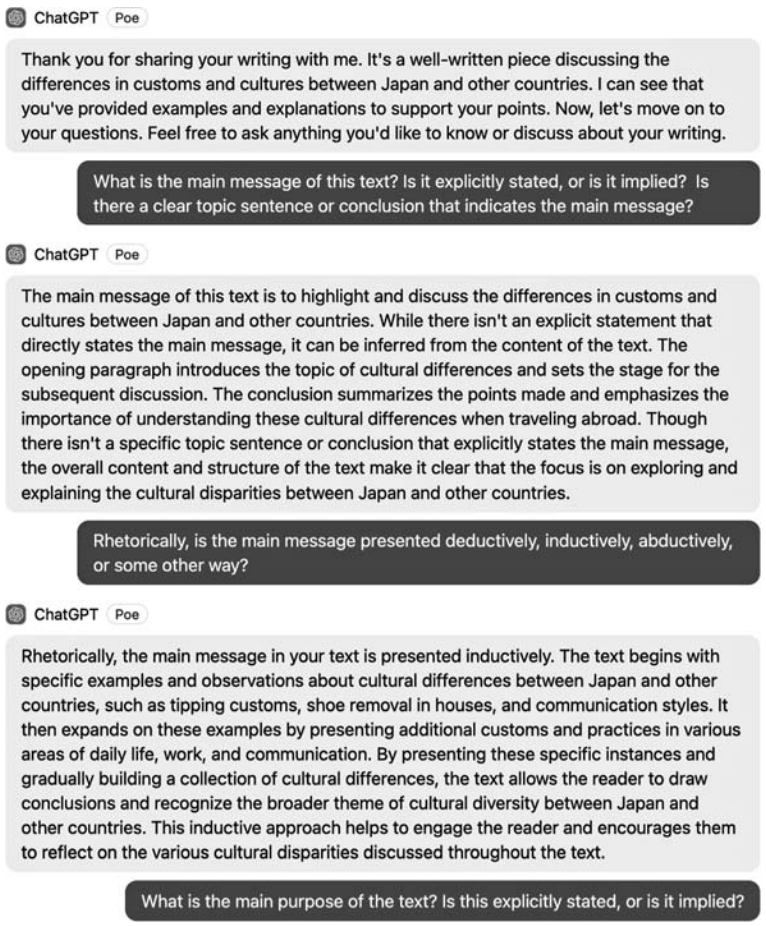


Figure 2 *AI Content Feedback Chat Example 1* for *S3*

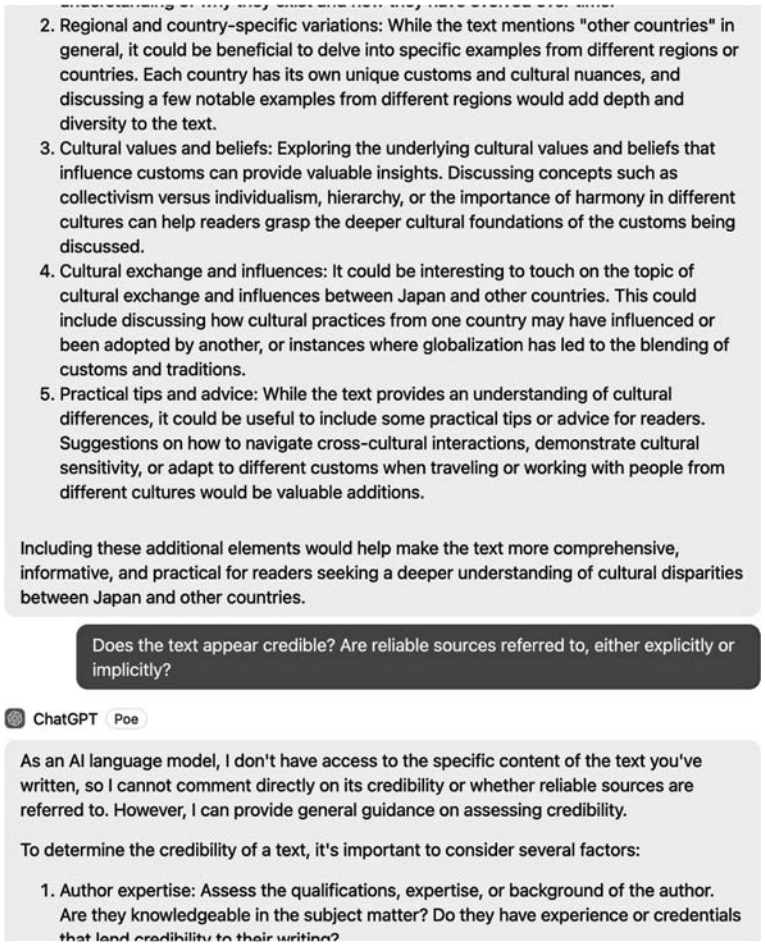


Figure 3 *AI Content Feedback Chat Example 2 for S3*
Note: Due to the length of the chat, the context window was apparently exceeded and feedback became generic advice when prompted to assess text credibility.

In addition to content-related prompts and questions, students were also given an example of how to receive language-related correction feedback in the form of a table indicating errors found, suggested corrections, and explanation of each error in Japanese (Figure 4). Students were told they could copy and paste each of these prompts and questions from the online resource. Excerpts from AI suggested language corrections are shown in Figures 5 and 6.

After getting writing feedback, start a new chat to do an English language check.

Initial prompt:

Act as an English teacher. I will enter the text of a presentation after you confirm the instructions. Please create a table with 3 columns. The first column should include each grammatical, word-choice, and spelling error you find in the text. The second column should list your suggested correction of the error you indicated. Content in the first 2 columns should be written in English. In the third column, for each error please include an explanation written in Japanese of why it was marked an error. OK? After you confirm these instructions, I will enter the text.

Enter text

" "

[Copy and paste your text + return. Do not include your name or other personal information.]

Figure 4 *Model Initial Chat Prompt for Obtaining AI Language Corrective Feedback*

Error	Correction	Explanation (説明)
four skills for improving English skills	four skills to improve English proficiency	"to improve" is more commonly used in this context.
I thought speaking is the most necessary in four skills.	I believe speaking is the most essential skill out of the four.	"believe" is more appropriate for expressing personal opinion.
if we can speak, we can accurately tell the other person what we want to say	if we can speak, we can effectively convey our thoughts to others.	"convey" is a more precise verb choice.
if we can pronounce properly, we will be able to listen without problems	if we can pronounce correctly, we will have better listening comprehension.	"correctly" is more suitable for describing pronunciation.
if we make mistakes, the others will correct those mistakes	if we make mistakes, others can help us correct them.	Revised for clarity and fluency.
it is important for us to experiences that talks with native speakers	it is important for us to have experiences of conversing with native speakers.	"experiences of conversing" is a more accurate phrasing.
If we have never talked with them in English, we must be nervous.	If we have never spoken to them in English, we are likely to feel nervous.	Revised for clarity and fluency.

Figure 5 *AI Language Correction Feedback for S2*

Note: AI produced explanations in English despite being prompted to display them in Japanese. AI may tend to default to the language of the submitted text.

Error	Correction	Explanation (Japanese)
pay a tip	leave a tip	チップを支払う
it is not customary to pay a tip	tipping is not customary	チップは習慣ではない
We have many customs and cultures in our lives	We encounter many customs and cultures in our lives	私たちは生活の中で多くの習慣や文化に触れます
when we enter a house in Japan	when entering a house in Japan	日本の家に入る時
it is normal to enter the house with shoes on	it is common to enter the house with shoes on	靴を履いたまま家に入るのが普通です
our commonplace customs and cultures may be different in other countries	customs and cultures in other countries may differ from ours	他の国の習慣や文化は私たちと異なる場合があります
First is daily life	First, let's talk about daily life	まず、日常生活について話しましょう
people often eat without saying anything	people often eat without any specific expressions	何も言わずに食事をすることが多い
people don't take a bath every day or only take a shower	people don't bathe every day or only take showers	毎日お風呂に入らない人もいるし、シャワーだけの人もいます
it is against manners to eat with dishes	it is considered impolite to eat with your hands	手で食事をすることはマナー違反です
overtime work is less because	overtime work is less	残業は少ないです。家

Figure 6 *AI Language Correction Feedback for S3*

Note: The excerpt includes examples of suggested language corrections that contain no grammatical or other errors but are potentially more statistically salient based on the training data.

The resource indicated which chatbots were likely to be most helpful for either content feedback or language feedback. This was based on previous trials by the instructor and the quality of feedback obtained. Students were told they were free to use any chatbot on Poe or elsewhere. They were also told that use of AI was optional. They were given time to practice in class and were encouraged to continue outside of class for the next session where they could present and discuss their results and actions. Between the conclusion of this class and the next session, the instructor also provided content-related feedback on texts submitted on the LMS using the review function of Microsoft Word. During the next session, students had an opportunity to discuss with peers and as a class their experience and the feedback they received. They then received the instructor's feedback and were asked to revise their texts for a final talk to be given in class and text submission on the LMS. Final talks were to be given in-class, without presentation slides or oral reading of a script. Assessment of the talk was holistic and based on both the final submitted text and delivery in class. For the revised text, the instructor confirmed for students that although the use of



Figure 7 *Item Frequency Results of Multiple-Choice Questions on Post-Task Questionnaire*

AI chatbot was optional, the sample questions regarding content and language feedback should be considered along with the instructor feedback.

Following the delivery of the talk in the subsequent class session, students were invited to answer an anonymous online survey (Figure 7) created with Microsoft Forms at the end of class regarding their perceptions of the use of chatbot for feedback on their talk texts as well as request for consent for their answers and submitted assignments to be analyzed and shared for research purposes by the instructor and author (confirmed by entering their name as signature). Students were told their participation in the research was voluntary and had no effect on their grade for the assignment or class and was done outside of class time, after the final class. One reminder invitation was sent electronically ten days after the last class. In total, six students responded to the questionnaire, and four consented to their submitted texts being analyzed and shared anonymously. Questionnaire results and frequency statistics were generated by Microsoft Forms. Further descriptive statistics were obtained using Microsoft Excel. Comparison of first draft that included instructor feedback and final revised texts was done using the document comparison function in Microsoft Word. The resulting comparison document was further analyzed along with student-submitted chatbot feedback chat history qualitatively in MAXQDA24.

4. Findings

Questionnaire results

Figure 7 displays the results for multiple-choice item questions (1, 3–10, 12, 14–23) in the post -task questionnaire, covering participants' prior experience using genAI, task scaffolding, feedback perceptions and uptake, future use, and overall satisfaction and problems encountered. Answers to short-answer items (2, 1, 13, 24) are described in the discussion that follows.

Trends & text feedback analyses

Prior experience (Questionnaire Q1–Q2)

Students had limited experience with genAI prior to this task. What experience they had was based on curiosity after hearing about it as a general topic and/or for some learning tasks related to classes, including this one. Uses included brainstorming and outlining for assignments, searching for information for completing assignments or understanding content, and non-specified chatting to "ask questions." Two students mentioned ChatGPT specifically, while others did not specify what kind of genAI they used. While not investigated here, given training data date limits and reliability issues related to current genAI, students will need to understand how genAI differs from search engines for finding information,

highlighting the need for AI literacy.

Task scaffolding implementation (Questionnaire Q3–Q7)

Student understanding of the task of using genAI for content and language corrective feedback on writing was mixed, with some students reporting good understanding and others less so. This again points to the need for both AI literacy and better task explanation and material. Since steps to be taken and questions to be asked were provided as task scaffolding, students reported little difficulty with the task. The majority of students (67%) reported using English only in completing the task, and remaining 33% using both languages. This was somewhat surprising since prompts were provided in both languages and the instructor indicated to students that using both languages would likely provide better results. It should be noted, however, that ChatGPT-3.5 appears to default to the language of the text entered (and not the language indicated in the prompt). Output language likely contributed to the later-reported difficulty in understanding the feedback generated in the task (Q10–Q13) by some students, which was not adjusted for length or students' proficiency level in the prompt. While chatting in the target language most certainly has language use practice benefits, for feedback purposes, first language is preferable if only one language is used. This also suggests a need for developing AI literacy, which includes purposes and methods of using tools as well as initial and follow-up prompting when chatbots do not output as expected. For the genAI chatbots used for both content and language feedback, the reported results were split evenly between ChatGPT-3.5 (free version 3.5 Turbo, January 2024) and Poe Assistant (free chatbot based on ChatGPT-3.5 Turbo, January 2024). For students who reported using Poe Assistant, it was not entirely clear if they were referring to the specific chatbot or confused it with the Poe site/app from which it is accessed. Interestingly, Poe Assistant itself was not listed as a recommended chatbot for the task, although students may have used it based on an understanding that it was based on ChatGPT. No other recommended chatbots were used by students participating in the study. This may have been due to the name recognition of ChatGPT. As was explained to students, certain genAI chatbots may be more helpful than others for certain tasks. This too falls under AI literacy and should be better highlighted in future activities and training.

Feedback perceptions and reported uptake (Questionnaire Q8–Q17)

Students used genAI chatbots to provide feedback on both the content, organization, rhetorical and other features of their talk draft texts as well as use of language with corrective suggestions and explanations. Overall, the results were mixed as well as interesting. For the feedback on content, all students reported that they found using genAI very useful. For language feedback, five out of six students reported using genAI for language feedback/correction as extremely or very useful, with one student reporting it less so for reasons unclear but possibly related to the amount of feedback/errors indicated (which

may have been overwhelming or possibly demotivating based on student comments). Compared to feedback provided by the instructor for content feedback, however, four out of six students (67%) reported it less useful, with two (33%) saying it was "about the same." Reasons given for perceiving it as less useful were because of too much output/feedback and difficulty in language (presumably in L2). One student said the instructor feedback was more targeted and succinct, making it easier to understand and use. Another student indicated a preference for human feedback as "reflective of a broader perspective." For language-focused feedback/correction, student perception of genAI fared better, with 83% (five out of six students) indicating that it was "about the same" as the instructor (despite the fact that no language feedback was given for language by the instructor). Students wrote that they found the quality to be reliable and cited the ability to easily get instant and useable feedback as benefits. However, one student found it less helpful than the instructor, citing the amount and difficulty of feedback (provided in L2). Together, results suggest that students found genAI useful but it was potentially qualitatively different from instructor feedback and depended on the kind of feedback—more useful for language-related corrective feedback and perhaps less so for content. In both cases, the amount of feedback was mentioned as well as difficulty of language. This points to an issue in the task design and prompts provided which did not indicate an appropriate length of feedback output for comprehension. In addition, it is clear that students would have likely found use of L1 (Japanese) for output in addition to or in place of L2 (English) useful in obtaining feedback. Both of these need to be considered in future task designs.

Although students did indicate genAI could be useful for generating feedback, their report of using that feedback is somewhat more complicated. For the content feedback, one student (17%) reported "a lot," three students (50%) "somewhat," and two (33%) "not very much." For the latter two students, this may have been because they preferred (as noted above) human feedback for content or found it to be difficult. Regarding the use of language correction feedback to correct their drafts, three students (50%) reported using it "not very much," one "somewhat," and two (33%) "a lot." Later examination of student drafts (see next section) confirmed this to be the case. Students who did incorporate the feedback did have higher quality final texts, although original drafts were of relatively higher quality as well.

Feedback & revision analyses (AI and instructor content and language feedback; student feedback uptake)

A total of 47 text content and organization feedback responses were generated by the AI chatbots versus 38 by the instructor (Table 2). As indicated in some of the student questionnaire responses, the length of AI-generated responses far exceeded the instructor's, with averages of 196 words vs. 13 words (also noted by Guo & Wang, 2023 and Su et al., 2023). The shortest AI response (81) was more than two-and-a-half times as much as the longest instructor's response (31). While greater amount of detailed feedback may have some

advantages, it can also become unmanageable from the learner's perspective and both the response length and level of language should be addressed in the prompt in future trials.

Table 2 *Student Text Content and Organization Feedback*

Student	AI feedback no. of responses	AI mean no. of words per reponse	AI no. of words per response range	Instructor feedbak commments no.	Instructor mean no. of words per comment	Instructor no. of words per response range
S1	16	174	81–325	7	10	1–20
S2*	–	–	–	9	14	3–28
S3	16	243	96–329	12	15	4–31
S4	15	170	71–298	10	13	2–23
Total	47	196	71–329	38	13	1–31

*AI feedback chat not submitted.

As an automated generative technology known for its limitations and inaccuracies ("hallucinations"), this analysis found the text content and organization feedback submitted by the three students to be a combined 96% wholly or mostly/largely accurate (Table 3). That is, the evaluative feedback was appropriate as rated by the researcher. This may also reflect findings in other research showing ChatGPT to be capable of relatively valid assessment (Baffour et al., 2024; Mizumoto & Eguchi, 2023; Pfau et al., 2023), based on the large model and specific data it has been trained on. However, since this assessment was completed by the researcher alone as a single-coder, these results are merely suggestive and await further research that involves more than one rater in order to better control reliability. Also, while it was found that accuracy was high for the text content and organization evaluative feedback, it often did not include suggestions for improvement, also mentioned by Su et al (2023). In an iterative process, learners will need to learn how to further prompt AI in order to obtain improvement suggestions as well. As indicated in the table, one student's chat exceeded the chatbot's context window and so subsequent feedback became general suggestions rather than anything specific to the submitted text (for this reason, it was not included in the analysis of accuracy). Students will also need to check for and deal with such issues.

Table 3 *AI Content Feedback Analysis*

Student	No. of feedback comments	Accurate	Largely or mostly accurate	Generally accurate
S1	25	19 (76%)	6 (24%)	0 (0%)
S2	–	–	–	–
S3	8*	3 (38%)	3 (38%)	2 (24%)
S4	13	12 (92%)	1 (8%)	0 (0%)
Total	46	34 (74%)	10 (22%)	2 (4%)

*Chat exceeded context window mid-way, rendering subsequent feedback as generic advice (non-specific to the student's text, although generally useful). Only text-specific feedback is counted.

In addition to evaluative feedback, several of the prompts provided to the students specifically requested suggestions for improvement of such things as the main text message, missing information, and general/overall items. As was the case for content and organization, the researcher assessed the quality of the suggestions in terms of helpfulness/usefulness. A total of 63% of the 16 suggestions in the texts were found to be helpful/useful or helpful/useful but needing more details or examples (Table 4). As a result, nearly 40% of the suggestions were judged to be only somewhat helpful and/or other (less than helpful, etc.). Compared to the content feedback, the quality was considerably lower, though better quality may be obtained with follow-up prompts by the learner.

Table 4 *AI Content Suggestions Analysis*

Student	No. of suggestions	Helpful/useful	Helpful/useful but needs details/ examples	(Only) somewhat helpful/other
S1	8 (5 unprompted)	4 (50%)	2 (25%)	2 (25%)
S2	–	–	–	–
S3	3*	1 (33.3%)	1 (33.3%)	1 (33.3%)
S4	5	2 (40%)	0	3 (60%)
Total	16	7 (44%)	3 (19%)	6 (37%)

Note: S2 did not submit AI content feedback chat.
*Chat exceeded context window mid-way, rendering subsequent suggestions as generic advice (non-specific to the student's text, although generally useful). Only text-specific suggestions are counted.

As for the content feedback and suggestions, the AI-suggested language corrective feedback submitted by two students was also assessed in terms of type and accuracy/quality. Overall, more than half of the overall corrective feedback related to collocations, or English vocabulary and expressions most often used in the specific context of the text and message based on the model's training data (Table 5). This meant that many of the suggested corrections were not necessarily errors in terms of grammar or even expression, but may have been less statistically likely based on the training data. Collocations were followed by style/register (19%) and word choice (14%) and there were noticeable differences

in these areas between students. Taken together, however, 85% of the corrections (S2: 76%; S 3: 93%) were largely lexical in nature, rather than grammatical, even though a number of grammatical errors were not flagged for correction by the AI (discussed below). Based on the very limited sample here, when given a general/broad prompt for corrective feedback that includes vocabulary, expressions, grammar, and mechanics, the AIs tended to focus on lexical suggestions common to English texts even if they were not necessarily errors, and appeared to ignore a number of grammatical errors (somewhat contradicting Wu et al., 2023). The reasons for this may include possible maximum response length limits as well as the nature of/biases in LLMs in general, but also text length and the limited number of cases (two) in the current analysis. In addition to the need for further research, instructors and learners will need to be aware of such issues in how to consider AI-generated feedback responses as well as how to confirm with further prompting.

Table 5 *AI-Suggested Language Corrections by Type and Student*

Type	S2	S3	Total
Collocation	10 (34.5%)	20 (69%)	30 (52%)
Style/register	9 (31%)	2 (7%)	11 (19%)
Word choice	3 (10%)	5 (17%)	8 (14%)
Verb tense/aspect/voice	3 (10%)	−	3 (5%)
Word form	1 (3.5%)	1 (3.5%)	2 (3%)
Singular/plural number	1 (3.5%)	1 (3.5%)	2 (3%)
Sentence/clause structure	1 (3.5%)	−	1 (2%)
Preposition	1 (3.5%)	−	1 (2%)
Total	29 (100%)	29 (100%)	58 (100%)

Text language corrective feedback was also assessed in terms of its accuracy by means of single-researcher evaluation. Based on this, the overall true positive rate was calculated as 74% (precision = .74), with false positives accounting for 26%, drawn from the lexical domain including collocations, style/register, and word choice (Table 6). As already indicated, AI tends to suggest corrections for items that are not necessarily errors, perhaps just less statistically salient based on its training data. All suggested grammar error corrections (100%) were determined to be true positives (though their smaller number may have contributed to this). Conversely, a total of 70% of the assessed false negatives (clear errors in the texts not singled out for correction by the AI) were grammar-related. The overall false-negative rate was a high 45% (recall = .55). Better and/or follow-up prompting may help in part to deal with this kind of issue.

Table 6 *AI-Suggested Language Correction Accuracy Analysis*

Errror/suggestion type	Positive freq.	True positive (TP) freq.	False positive (FP) freq.	False negative (FN) freq.	Precision (%)*	Recall (%)**
Collocation	30	23	7	8	77%	74%
Style/register	11	7	4	−	64%	100%
Word choice	8	5	3	−	63%	100%
Verb tense/aspect/voice	3	3	0	2	100%	60%
Word form	2	2	0	−	100%	100%
Singular/plural number	2	2	0	5	100%	29%
Sentence/clause structure	1	1	0	3	100%	25%
Preposition	1	1	0	6	100%	14%
Punctuation	0	−	−	7	−	0%
Article/determiner	0	−	−	3	−	0%
Coherence break	0	−	−	2	−	0%
Total	58	44	14	36	76%	55%

*precision = TP/(TP + FP); **recall = TP/(TP + FN)

Lastly, student uptake of AI-generated feedback in their revised texts was analyzed in comparison to the instructor feedback. All students who submitted their revised texts for analysis increased in the total number of words they wrote, with an average of 96 words (Table 7). No clear uptake was confidently determined from the AI-generated content feedback in the revised texts, however, although as already indicated, some students replied in the questionnaire that they incorporated some AI feedback. It is possible that there is a gap between what was reported and what really took place, but more likely AI feedback and instructor feedback overlapped to some extent and students opted to revise their texts in the places indicated by the instructor. Recall that several of the students replied that they used very little (if any) of the AI content feedback due to either a preference for human feedback as being broader and more nuanced, or to difficulty in comprehending output in L2. Overall, learners were judged by the researcher to have incorporated 76% of the instructor content-related feedback. As for AI language corrective feedback suggestions, two students (S1 and S4) did not submit their feedback but indicated that they incorporated little of it into their revised texts. This was reflected in the language revisions assessed by the researcher, with just a small number of unknown origin for each since the instructor did not include any language feedback in either text. For the other two students who submitted the AI language corrective feedback, however, there were significant (though varying) revisions based on it, with S2 incorporating 93% of it as uptake, while S3 was a much smaller 41%. S3 was much more selective in using the feedback which was 93% composed of collocation, style/register, and word choice suggestions.

Table 7 *Student Text Revisions by Kind and Source*

Student	Draft no. of words	Revision no. of words	Word increase	Instructor no. of feedback comments	Student no. of content revisions	AI no. of sugg. language corrections	Student no. of lang. revisions	Student total no. of revisions	Student no. of non-revisions
S1	344	490	146	7	6	-	3	9	1
S2	256	329	73	9	8	29	29*	29	2**
S3	656	778	122	12	7	29	19*	26	15**
S4	247	288	41	10	8	-	2	10	3
Mean	376	471	96	10	7	29	13	19	5

Notes:
1. There was no clear content revision based on AI feedback, though some feedback overlapped with instructor feedback. Total uptake of instructor feedback: 29/38＝76%.
2. S2 did not submit AI content feedback; S3's AI chat exceeded the context window mid-chat, causing subsequent feedback to become generic advice (non-specific to the originally submitted text).
3. S1 and S4 did not submit AI-suggested language revision results.

*S2: 27 language revisions based on AI feedback (＝93% uptake), 2 of indeterminate origin; S3: 12 language revisions based on AI feedback (＝41% uptake); 2 revisions based on instructor feedback; 5 revisions of indeterminate origin. S 1 and S2 language revisions were of indeterminate origins. Total uptake of AI-suggested language corrections: 39/58 ＝67%.

**S2: One AI and one instructor langauge feedback non-revision; S3: 14 AI and one instructor language feedback non-revisions.

Issues of how best to generate feedback, for what, and how to use the results of that are all areas for future investigations but all may be included under the topic of AI literacy for language learning. Student enjoyment of using genAI for feedback (questionnaire items 16 and 17) depended on the learner and was relatively evenly divided, with a slight preference for content feedback rather than language. It is possible that generative language correction feedback, with its form- rather than content-focus may have been less interesting for one student.

Future use (Questionnaire Q18–Q21)

Students were generally positive toward using genAI again in the future to generate both content and language correction feedback on their English writing, with a preference for the latter (67% vs. 84%). Four out of six students (67%) indicated they would very likely use it again for feedback, with two (33%) neutral/unsure. All students indicated they wanted to use genAI for other (unspecified) purposes.

Task satisfaction and problems (Questionnaire Q22–Q24)

Regarding final questions about learning task satisfaction and problems encountered, all students reported being "satisfied" (4 out of 5 on Likert scale), with few (50%) or virtually no (50%) problems. Despite issues already indicated above, it appeared task scaffolding was acceptable in terms of students' experience of the task. However, some students did report a gap between their intentions/expectations and the results they obtained. More experience, greater proficiency in using, and overall AI literacy should help to address such issues.

5. Discussion & Conclusion

Research questions and implications

Based on the data and analysis presented above, tentative answers to the research questions are as follows:

1. How can genAI be used as a feedback tool to support English language learners' process writing?

The findings suggest that at present, at least for using well-known and freely available chatbots such as ChatGPT-3.5, genAI is best used as a supplementary feedback tool. Based on the analysis here, there appears to be potential for aiding in content and language development if used appropriately and judiciously. It should be used alongside human (instructor) feedback, with appropriate and iterative prompting, controlled output length, and awareness of tendencies such as some lack of specificity, language/lexical corrective biases toward vocabulary, collocations, and expressions, and a potential for both over- and under-correction. As a result, AI literacy training is essential. Additional use of specialized and freely available corrective tools such as Grammarly, as well as automated writing evaluation programs, make genAI chatbots one among several writing resources available to learners as described by Mizumoto (2023).

2. What prior experience do students have using genAI?

Based on self-reported responses and researcher examination of submitted materials, it appears students participating in the study had limited prior experience with using genAI for writing feedback or other purposes. However, this will likely change due to both increasing development and promise and likely prominence of genAI in the future. Focusing efforts to appropriately incorporate it into education remains an urgent task.

3. How do students use genAI as a feedback tool for checking their writing?

There appeared to be a preference for ChatGPT specifically or as the base of the Poe Assistant chatbot, despite the additional recommendations in Poe (Claude for content and Gemini for language feedback). This likely had to do with the name recognition or perhaps the default chatbot in the Poe application. Some prior familiarity may also have been possible. For prompting, students used the resource prompts exactly as provided through copying and pasting. Students did not appear to add their own follow-up prompts for clarification, translation, etc. but accepted the output as provided by the chatbot. Since it appears that the chatbots had a tendency to output in English even when prompted in Japanese (after an English text was entered) and students replied that the length and ease of understanding were problems, learning iterative, follow-up prompting will become important for effective

use. Effective prompt engineering will need to be emphasized in training and developing AI literacy.

4. How do they perceive the process and results of using genAI as a feedback tool for checking content, organization, and language in their writing?

While the process did not appear to cause difficulties for students and overall they were positive about their use of genAI for feedback on their writing and for continued future use, some students indicated difficulties in understanding and using the output. Although responses were not uniform, there was a clear preference for language compared to content feedback. This is consistent with Ranalli (2021) and others (Alharbi, 2023) in research with Grammarly and other automated tools, finding that discrete language feedback was preferred due to its ease of understanding and implementation. The students in the present study generally rated the targeted and more succinct instructor content feedback as more helpful than the AI feedback, with the language feedback being roughly equal. Despite this preference, given the nature of genAI, it can be a potentially unreliable tool without an understanding of how it works (e.g., suggesting alternatives even when there are no errors or the possibility of missing grammatical and other errors). Instruction in appropriate use and experience will be necessary for continued use (in conjunction with and support of instructor feedback).

5. How accurate and useful are the writing feedback results provided by genAI and are they similar to or different from human/instructor feedback?

In the present study, accuracy and usefulness depended on the type of feedback, as well as the areas within each type (based on the analysis of a single researcher/coder). Content feedback was found to be 96% accurate or mostly accurate, a quite high level. Suggestions for content were a lower 63% helpful/useful with and without additional explanation. For language feedback, genAI had a tendency to focus on lexical suggestions rather than grammar corrections. While precision was a relatively impressive .74, recall was just .55, indicating a relatively large number of missed errors, 70% of which were grammar-related. The student reported preference for language feedback and the possibility for genAI to focus more on collocation/lexical suggestions versus grammatical error correction needs to be addressed and understood by both instructors and learners. It should be noted, however, that variables such as text length, proficiency level, as well as prompting and specific chatbot and version used may affect results. Overall, these results appear to generally be consistent with those reported in Baffour et al. (2024), Guo and Wang (2023), Mizumoto and Eguchi (2023), Pfau et al. (2023), Su et al. (2023), and Wu et al. (2023) regarding ChatGPT's assessment and correction capabilities (while partial differences with Wu et al. may be accounted for by variable differences just mentioned).

6. How do students use the results obtained from genAI as a feedback tool for checking their writing?

Results from the survey and text analysis found students used AI feedback differentially and selectively. Based on student responses, there were differing levels of genAI feedback preference and use (consistent with Escalante et al., 2023). Some clearly preferred instructor to AI for content feedback, with one student suggesting that the latter lacked the kind of perspective that humans have. Others found the language feedback lacking, and even some students who found it potentially helpful made use of it selectively. Just as genAI is tied to individualized learning, it also depends on the individual and how they use and respond to it, making use an individual experience. In this regard, Ranalli's (2021) concept of "calibrated trust" in his study of students' differential use of Grammarly is informative.

In summary and by implication, the findings in the present study suggest the following:

1. Use of AI for writing feedback purposes needs to be in addition to, not in place of, instructor feedback.
2. Developing AI literacy will be important for successful use. It does not necessarily develop on its own and appropriate task design, scaffolding, and repeated experience will all contribute positively.
3. AI literacy includes understanding potential benefits, problems, and biases in using, along with the types of tools available and when and how to use them (Warschauer et al., 2023). In particular, choosing appropriate tools, and prompting them to provide comprehensible and useable feedback in terms of language, amount, etc. should be emphasized.
4. Use of L1 is almost certainly more helpful for feedback purposes.
5. For content feedback, asking a series of questions one by one will provide more detailed and potentially useful feedback rather than entering all questions/feedback criteria all at once. Alternatively, prompt engineering may achieve similar results and should be explored further. While more detailed feedback may be more useable, there should be limits on its amount and complexity to avoid overwhelming learners.
6. Appropriate output in terms of length and difficulty should be included in the prompt.
7. Possible tendencies in the chatbot model (GPT-3.5) employed here to focus on collocations and lexical expressions compared to grammar correction should be both understood by users and researched further.
8. Other issues such as feedback quality/validity and reliability need to be further explored.
9. There is potential for genAI feedback in a broader feedback and learning regimen that can foster greater learner autonomy, self-efficacy, instructor efficacy, and overall metacognitive development and learning.

Limitations & future directions

This was an exploratory study with a very small sample size. In addition, genAI feedback output and student revised text data analyses were undertaken based on a single coder, and a number of variables such as AI content feedback uptake, text length, and learner proficiency could not be accounted or controlled for. Finally, only a single chatbot/model, ChatGPT-3.5 was employed, and different genAI and versions may obtain different results. As a result, the study can not claim generalizability or confirmed reliability. Findings can only be taken as suggestive, informing future investigations and cautious classroom practice. Future research and practice should improve on task-design and direction, training, and measurement and employ more sophisticated research methods, both quantitative and qualitative. Given the likely continued development and spread of generative and other AI technology, it is crucial that both pedagogy and research address the affordances and significant challenges that it presents second language teaching and learning.

References

Alharbi, W. (2023). AI in the foreign language classroom: A pedagogical overview of automated writing assistance tools. *Education Research International*, *2023*, 1-15. https://doi.org/10.1155/2023/4253331

Baffour, P., Saxberg, T., Abboud, R., Boser, U., & Crossley, S. (2024). *Assessing ChatGPT's writing evaluation skills using benchmark data*. The Learning Agency. https://the-learning-agency.com/guides-resources/assessing-chatgpt-writing-evaluation-skills-using-benchmark-data/

Barrot, J. S. (2023a). ChatGPT as a language learning tool: An emerging technology report. *Technology, Knowledge and Learning, 29*, 1151-1156. https://doi.org/10.1007/s10758-023-09711-4

Barrot, J. S. (2023b). Using ChatGPT for second language writing: Pitfalls and potentials. *Assessing Writing*, *57*, 100745. https://doi.org/10.1016/j.asw.2023.100745

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. arXiv. https://doi.org/10.48550/arXiv.2005.14165

Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2023). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*, *61*(2), 228-239. https://www.tandfonline.com/doi/full/10.1080/14703297.2023.2190148

Dzekoe, R. (2017). Computer-based multimodal composing activities, self-revision, and L2 acquisition through writing. *Language Learning and Technology*, *21*(2), 73-95. https://dx.doi.org/10125/44612

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. https://doi.org/10.1186/s41239-023-00425-2

Fütterer, T., Fischer, C., Alekseeva, A., Chen, X., Tate, T., Warschauer, M., & Gerjets, P. (2023). ChatGPT in education: Global reactions to AI innovations. *Scientific Reports*, *13*(1), 15310. https://doi.org/10.1038/s41598-023-42227-6

Godwin-Jones, R. (2022). Partnering with AI: Intelligent writing assistance and instructed language learning. *Language Learning and Technology*, *26*(2), 5-24. http://doi.org/10125/73474

Godwin-Jones, R. (2023). Emerging spaces for language learning: AI bots, ambient intelligence, and the metaverse. *Language Learning and Technology*, *27*(2), 6-27. https://hdl.handle.net/10125/73501

Guo, K., & Wang, D. (2023). To resist it or to embrace it? Examining ChatGPT's potential to support teacher feedback in EFL writing. *Education and Information Technologies*, *29*, 8435-8463. https://doi.org/10.1007/s10639-023-12146-0

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., ... Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kong, S.-C., Lee, J. C.-K., & Tsang, O. (2024). A pedagogical design for self-regulated learning in academic writing using text-based generative artificial intelligence tools: 6-P pedagogy of plan, prompt, preview, produce, peer-review, portfolio-tracking. *Research and Practice in Technology Enhanced Learning*, *19*, 030. https://doi.org/10.58459/rptel.2024.19030

Lampropoulos, G., Ferdig, R. E., & Kaplan-Rakowski, R. (2023). A Social media data analysis of general and educational use of ChatGPT: Understanding emotional educators. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4468181

Liang, W., Zhang, Y., Cao, H., Wang, B., Ding, D., Yang, X., Vodrahalli, K., He, S., Smith, D., Yin, Y., McFarland, D., & Zou, J. (2023). Can large language models provide useful feedback on research papers? A large-scale empirical analysis. arXiv. https://doi.org/10.48550/arXiv.2310.01783

Mizumoto, A. (2023). Data-driven learning meets generative AI: Introducing the Framework of Metacognitive Resource Use. *Applied Corpus Linguistics*, *3*(3), 100074. https://doi.org/10.1016/j.acorp.2023.100074

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, *2*(2), 100050. https://doi.org/10.1016/j.rmal.2023.100050

Moorhouse, B. L., Yeo, M. A., & Wan, Y. (2023). Generative AI tools and assessment: Guidelines of the world's top-ranking universities. *Computers and Education Open*, *5*, 100151. https://doi.org/10.1016/j.caeo.2023.100151

Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. *Computers and Education*: *Artificial Intelligence*, *2*, 100041. https://doi.org/10.1016/j.caeai.2021.100041

OpenAI (n.d.-a). *GPT-4*. Retrieved September 16, 2024, from https://openai.com/gpt-4

OpenAI (n.d.-b). *OpenAI Models*. Retrieved September 16, 2024, from https://platform.openai.com/docs/models

OpenAI (2020). *OpenAI API*. Retrieved September 16, 2024, from https://openai.com/index/openai-api/

OpenAI (2022). *Introducing ChatGPT*. Retrieved September 16, 2024, from https://openai.com/index/chatgpt/

OpenAI (2024). *Hello GPT-4o*. Retrieved September 16, 2024, from https://openai.com/index/hello-gpt-4o/

Pack, A., & Maloney, J. (2023). Potential affordances of generative AI in language education: Demonstrations and an evaluative framework. *Teaching English With Technology*, *23*(2). 4-24, https://doi.org/10.56297/BUKA4060/VRRO1747

Perkins, M. (2023). Academic integrity considerations of AI large language models in the post-pandemic era: ChatGPT and beyond. *Journal of University Teaching & Learning Practice*, *20*(2). https://doi.org/10.53761/1.20.02.07

Pfau, A., Polio, C., & Xu, Y. (2023). Exploring the potential of ChatGPT in assessing L2 writing accuracy for

research purposes. *Research Methods in Applied Linguistics*, *2*(3), 100083. https://doi.org/10.1016/j.rmal.2023.100083

Quora, Inc. (n.d.). *Poe*. https://poe.com/

Ranalli, J. (2021). L2 student engagement with automated feedback on writing: Potential for learning and issues of trust. *Journal of Second Language Writing*, *52*, 100816. https://doi.org/10.1016/j.jslw.2021.100816

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

Su, Y., Lin, Y., & Lai, C. (2023). Collaborating with ChatGPT in argumentative writing classrooms. *Assessing Writing*, *57*,100752. https://doi.org/10.1016/j.asw.2023.100752

Sullivan, M., Kelly, A., & McLaughlan, P. (2023). ChatGPT in higher education: Considerations for academic integrity and student learning. *Journal of Applied Learning and Teaching*, *6*(1), 31-40. https://doi.org/10.37074/jalt.2023.6.1.17

Vox (2023, December 12). *AI can do your homework. Now what?* https://youtu.be/bEJ0_TVXh-I?si=UdDowUlG8oibyTeN

Walters, W. H., & Wilder, E. I. (2023). Fabrication and errors in the bibliographic citations generated by ChatGPT. *Scientific Reports*, *13*, 14045. https://doi.org/10.1038/s41598-023-41032-5

Warschauer, M., Tseng, W., Yim, S., Webster, T., Jacob, S., Du, Q., & Tate, T. (2023). The affordances and contradictions of AI-generated text for second language writers. *Journal of Second Language Writing*, *62*, 101071. https://doi.org/10.1016/j.jslw.2023.101071

White, S. A. (2023). Developing critical thinking skills in virtual exchange: Reflective journal writing for university English as a lingua franca learners. *The Ryukoku Journal of Humanities and Sciences*, *45*(1), 39-55. http://doi.org/10.50873/10773

Wu, H., Wang, W., Wan, Y., Jiao, W., & Lyu, M. (2023). ChatGPT or Grammarly? Evaluating ChatGPT on Grammatical Error Correction Benchmark. arXiv. https://doi.org/10.48550/arXiv.2303.13648

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., ... Wen, J.-R. (2023). A survey of large language models. arXiv. https://doi.org/10.48550/arXiv.2303.18223