https://doi.org/10.29140/9780648184485-09

Using Large Language Models (LLMs) to Facilitate L2 Proficiency Development Through Personalized Feedback and Scaffolding: An Empirical Study

Naiyi Xie Fincham[1,*] and Aitor Arronte Alvarez[1]
[1]University of Hawaiʻi at Mānoa, USA

Abstract

This present research is part of a longitudinal study that focuses on understanding the capabilities and affordances of generative artificial intelligence (AI), especially Large Language Models (LLMs), and how they can be harnessed to support second language (L2) learning, teaching, and research. Leverating LLMs' significant new capabilities that resemble those of human languages, an LLM-based application is developed to facilitate English language learners' proficiency growth in productive skills. Specifically, utilizing the GPT-4o model by OpenAI, this application guides learners through implicitly constrained text-based conversations based on learners' current proficiency level and personal interests. Further, two individualized feedback mechanisms provided by the AI drawing on learner output are programmed and embedded in the conversation practices: 1) graduated corrective feedback informed by the sociocultural theory of learning to guide learners' self-identification and self-correction of errors, and 2) metacognitive reviews to help learners become aware of areas for improvement according to the detailed proficiency descriptions outlined in the Common European Framework of Reference for Languages (CEFR). The research presented here centers on the current phase in this project that aims to evaluate the quality and effectiveness of the feedback mechanisms, examine learners' responsiveness to AI-generated graduated feedback, and assess the system's potentials for supporting independent language learning, classroom integration, and second language (L2) research. The main data sources include

*Corresponding Author E-mail: naiyixf@hawaii.edu

In Y. Wang & M. S. Cárdenas Claros (Eds.), Proceedings of the XXIInd International CALL Research Conference, 2024 (pp. 59–64). Castledown.

conversation logs between seven English learners at different proficiency levels and the AI system over four weeks, and learners' responses to post-session evaluation surveys. Findings from this phase provide important insights into how LLMs can be utilized to guide conversation practices and provide both personalized corrective and metacognitive feedback and scaffolding for language learners. Results emerged from the data analysis so far have informed further modifications to the feedback mechanisms to enhance their effectiveness and overall learner experience and led to pedagogical implications for classroom instruction and independent learning.

Keywords: Dialogue-based CALL, Large language models (LLMs), L2 productive skills, Corrective feedback, Learner-AI conversation

Background

This present research is part of a longitudinal study that focuses on understanding the capabilities and affordances of generative artificial intelligence (AI), especially Large Language Models (LLMs) and how they can be harnessed to support second language (L2) learning, teaching, and research. Recent developments in AI have brought significant changes to language modeling; as a result, LLMs have acquired significant new capabilities that resemble those of human languages: zero-shot language generation, in-context learning, and chain of thought prompting which elicits reasoning (Liu et al., 2023). Leveraging these capabilities, an LLM-based (GPT-4o by OpenAI) application is developed to facilitate English language learners' proficiency development in productive skills through simulated conversation practices. Drawing on previous developments and research on dialogue-based Computer-Assisted Language Learning (CALL) systems (Bibauw et al., 2019), this application is designed to guide learners through implicitly constrained text-based conversations customized to learners' personal interests and proficiency levels informed by the CEFR (Council of Europe, 2020). Further, two individualized feedback mechanisms provided by AI based on learner output are programmed on top of the same GPT model and embedded in the conversation practices: 1) graduated corrective feedback informed by the sociocultural theory of learning to guide learners through self-identification and self-correction of errors (Vygotsky, 1987), and 2) metacognitive summary reviews to enhance learners' planning, self-monitoring, and self-reflection of their own learning processes and products (Sato, 2022).

A graduated approach to corrective feedback is driven by each individual learner's real-time responses to potential errors, which may promote L2 learning in a dialogically and collaboratively constructed zone of proximal development (ZPD) (Aljaafreh & Lantolf, 1994). The feedback progresses from most general and implicit (e.g., asking learners to reexamine a previous sentence), to most specific and explicit (e.g., directly highlighting the error with metalinguistic explanations) that help learners self-identify and self-correct errors in their own output (e.g., Ai, 2017; Poehner & Lantolf, 2013). In the case of this AI-powered application, such graduated feedback and scaffolding are implemented via iterative prompting (Wang et al., 2022). After each turn by the participant, a prompt mechanism is carried out to check whether the participant's response contains any error (grammatical, syntactic, or semantic). If correction is required, a process consisting of three mediating levels of explicitness and hints engages the learners in self-identification and correction of errors until the correct form is provided.

For the metacognitive summary reviews, the system is instructed to draw on the detailed description of goals and objectives in the CEFR proficiency guidelines and each learner's output during the practice session, and to present evaluative summaries and suggestions for improvement. Metacognitive reviews are delivered at two separate timepoints for each practice session: 1) a midpoint review after a few completed turns and 2) an extended summary after the session is completed.

The research presented in this year's International CALL conference focuses on the current phase in this project that aims at evaluating the quality and effectiveness of the graduated feedback mechanisms, examining the characteristics of learner-AI conversational interaction (especially learners' responsiveness to AI-generated feedback), and assessing the system's potentials for supporting independent language learning, classroom integration, and L2 research.

Overview of Methodology

Data Collection

A total of seven participants were selected for this phase of the research. The participants engaged in extended text-based conversations with the AI agent at the proficiency levels selected by themselves and were instructed to practice at least once a week for a month at their own convenience. At the time of this report, this has resulted in 26 conversations, 91 adjacency pairs, and 182 turns, with proficiency distributions of 3.9% for A1, 23.1% for B1, 53.8% for C1, and 19.2% for C2 on the CEFR scale. Interaction transcripts between participants and the AI agent are extracted from the system and converted into text files for analysis. At the end of each conversation practice session, participants were prompted to complete a post-session evaluation survey (exit survey) to provide their rating and feedback to determine: 1) how the system guided their language production, 2) helpfulness of the grammar feedback, 3) helpfulness of the feedback to communication skills, and 4) the degree to which they attended to AI-generated target language.

Data Analysis

First, interaction transcripts between learners and the AI agent for all the conversation sessions are analyzed qualitatively to explore participants' moment-to-moment responsiveness and changes in reaction to the AI-generated questions and prompts. Second, responses from the post-session evaluation survey are analyzed both qualitatively and quantitatively to assess the efficacy of the two feedback and scaffolding mechanisms from the learners' perspective.

Preliminary Results

While the data analysis is still ongoing, some preliminary results have emerged. First, participants were able to engage in extended conversations with the AI agent and were attentive to the AI agent's prompts to identify and correct their own errors, with varying degrees and patterns identified in their responsiveness. Second, the exit survey responses show overall positive ratings provided by the participants regarding the quality and helpfulness of the AI-generated feedback on their overall output, grammar, and communication skills. Further,

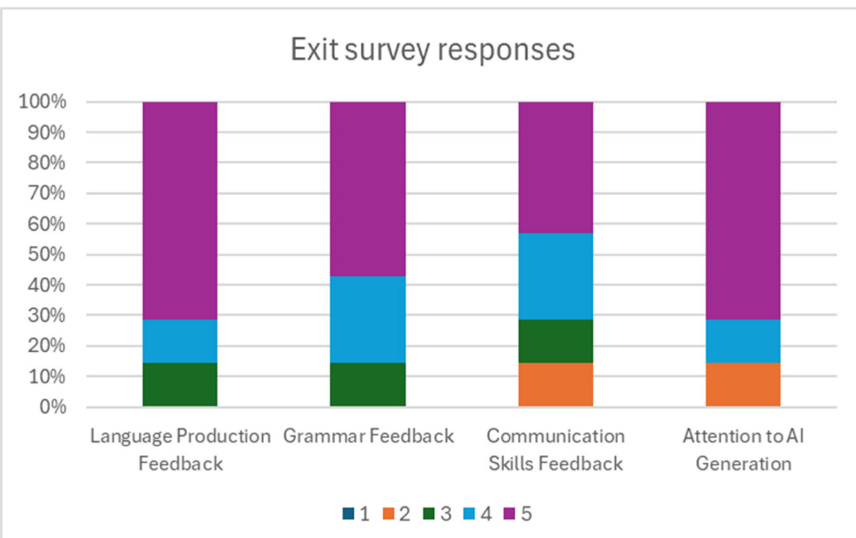

Figure 1. Summary of participants' ratings to exit survey questions.

the results support the observations drawn from the interaction data that participants would pay attention to the target language generated by the AI agent and consider it as a useful learning resource (see Figure 1). Meanwhile, some participants have shared less favorable evaluations on the effectiveness of the graduated approach to corrective feedback, citing its intrusiveness on the conversational flow and excessive focus on form over meaning.

Discussion and Conclusions

Findings from this phase of the study indicate that the LLM-based conversational AI system can provide personalized feedback and scaffolding drawing on individual learner's target language production, which is a significant improvement on one of the major limitations related to natural language understanding presented in previous dialogue-based CALL systems (Bibauw et al., 2019). However, participants' mixed reactions to the error correction experience guided by the graduated feedback imply that the feedback mechanism needs to be further fine-tuned through prompting instructions. Further, learners' positive perceptions regarding the individualized holistic language experience and the metacognitive review mediated by established proficiency guidelines show that a system as such can both benefit independent language learning and support classroom instruction with appropriate instructional design.

References

Ai, H. (2017). Providing graduated corrective feedback in an intelligent computer-assisted language learning environment. *ReCALL*, *29*(3), 313–334. https://doi.org/10.1017/S095834401700012X

Aljaafreh, A., & Lantolf, J. P. (1994). Negative feedback as regulation and second language learning in the zone of proximal development. *The Modern Language Journal*, *78*, 465–483. https://doi.org/10.1111/j.1540-4781.1994.tb02064

Bibauw, S., François, T., & Desmet, P. (2019). Discussing with a computer to practice a foreign language: Research synthesis and conceptual framework of dialogue-based CALL, *Computer Assisted Language Learning*, *32*(8), 827–877. https://doi.org/10.1080/09588221.2018.1535508

Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume*, Council of Europe Publishing, Strasbourg. www.coe.int/lang-cefr.

OpenAI. (2024). *Hello GPT-4o*. Retrieved August 20, 2024, from https://openai.com/index/hello-gpt-4o/

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, *55*(9), 1–35.

Poehner, M. E., & Lantolf, J. P. (2013). Bringing the ZPD into the equation: Capturing L2 development during computerized dynamic assessment (C-DA). *Language Teaching Research*, *17*(3), 323–342. https://doi.org/10.1177/1362168813482935

Sato, M. (2022). Metacognition. In *The Routledge handbook of second language acquisition and individual differences* (pp. 95–110). Routledge.

Vygotsky, L. S. (1978). *Mind in society*. Cambridge, MA: Harvard University Press.

Wang, B., Deng, X., & Sun, H. (2022). Iteratively Prompt Pre-Trained Language Models for Chain of Thought. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 2714–2730).