PeerChat: Scaling Peer Tutoring with AI Tutees

Ojas Karnavat[1], Soren Rosier[2]
[1]University of California, Berkeley
[2]Stanford University
ojask@berkeley.edu, rosier@stanford.edu

Abstract

This paper explores the innovative use of generative artificial intelligence (AI) as a tutee in peer tutoring sessions, addressing limitations in scaling peer tutoring. Leveraging OpenAI's ChatGPT models, we introduce PeerChat, an AI tutee designed to facilitate student-led tutoring interactions. Peer tutoring, known to benefit both tutors and tutees, is limited to times when a teacher decides to incorporate peer tutoring into their lesson plan and thereby carries high stakes since the tutor is responsible for another student's learning. By employing generative AI to build a realistic AI tutee in one-on-one peer tutoring sessions, students gain opportunities for independent teaching practice, addressing these limitations. Our study, the first of its kind, systematically investigates the creation of an AI tutee and evaluates its performance against predefined criteria. Results indicate promising accuracy, with successive models improving response quality. This novel approach holds the potential to enhance peer tutoring experiences and opens the door to further research in optimizing AI tutees for education.

Introduction

The advent of generative artificial intelligence (AI) has great potential for impact in education. One common focus of utilizing AI in classrooms has been through creating AI tutors; a decades-old example is Cognitive Tutor (Carnegie Learning, 2010). Indeed, the impact of 1-on-1 tutoring by teachers and professional tutors is unrivaled (Lepper and Woolverton, 2002), so efforts to replicate similar success with AI could help scale personalized learning by making 1-on-1 tutoring affordable and accessible to a wider audience. However, the inverse—an *AI tutee*—is a relatively untapped utilization of AI that shows promise in solving primary limitations of peer tutoring. Peer tutoring in classrooms has been shown to increase learning not just for tutees but also for the tutors (Bowman-Perrott et al., 2013; Leung, 2015; Kobayashi, 2019). Simulated tutees have also been shown to demonstrate the protégé effect, highlighting another benefit of building AI tutees (Chase et al., 2009). Creating realistic AI tutees that students can teach will help address three important limitations of scaling peer tutoring:

- Peer tutoring is logistically challenging, so students are only able to gain the benefits of being a tutor when it is in accordance with the teacher's lesson plan.
- Students rarely have opportunities to practice teaching and become better teachers outside of the classroom.
- The opportunity to practice teaching in a safe environment (outside a real tutoring session) could boost both confidence and ability to do real teaching.

Using generative AI to simulate a tutee in a 1-on-1 peer tutoring session allows students to practice teaching outside the classroom at their own pace, thereby making the benefits of teaching more accessible. This aligns with Taylor's (1980) proposal that computers can play a role in education as tutors, tools, and tutees. While the use of computers as a tool and tutor is common, utilizing computers as a tutee is less explored (Tate et al., 2023). There have been efforts such as SimStudent where a bot is taught a task with the goal of helping the bot learn that task, but such systems are not focused on helping a real student grow (Matsuda, Cohen, and Koedinger, 2014). The closest work to our paper is GPTeach, which helps aspiring teachers practice tutoring with simulated tutees (Markel et al., 2023). While GPTeach also aims to build an AI tutee, it enables a different use case and does not investigate best practices for such a system by comparing different models. The premise of this paper is that since students learn by teaching, giving them more chances to teach can help them become more proficient on the topic and become better teachers, which can translate into a better peer tutoring session with a real student.

Therefore, we propose PeerChat—an AI tutee that a student can teach. PeerChat is built from ChatGPT (OpenAI). In this paper, we aim to understand how much the output of this system resembles how a real student might respond, which is crucial to advance the role of computers as a tutee (Polverini and Gregorcic, 2023). This study is, to our knowledge, the first systematic investigation of building an AI tutee using various models of ChatGPT and evaluating each model on predefined benchmarks.

Methodology

In order to build and test this system, the following steps were followed:

1. The AI tutee (PeerChat) was built through prompt engineering using ChatGPT models.
2. An interactive module was created (Figure 1) that allows users to tutor PeerChat by entering text in their dialogue box, to which PeerChat responds.
3. 300 interactions with PeerChat were scored on four evaluation criteria to understand how well PeerChat simulates a middle school student.

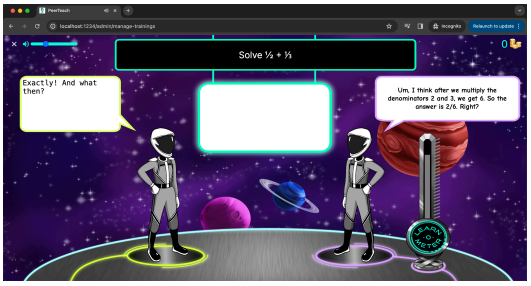

Fig. 1. An interactive module to test PeerChat

Zero-shot prompting expectedly did not perform well, likely since it is difficult to meet the criteria below without successful examples. Another popular approach, chain of thought prompting (Wei et al., 2022), also underperformed since it is meant for complex reasoning tasks as opposed to roleplaying a particular character. Our most successful model used few-shot prompting (Brown et al., 2020) with 3 examples, the results of which are shown in this paper.

We tested PeerChat with 3 ChatGPT models (Table 1):

Model Name	Context Window	Release Date
gpt-4-1106 preview	128,000 tokens	Nov 6, 2023
gpt-4	8,192 tokens	June 13, 2023
gpt-3.5-turbo	4,096 tokens	June 13, 2023

Table 1: ChatGPT Models used for PeerChat

Each model aimed to resemble an AI tutee for the problem, "Add one-half and one-third," and was intended to behave as a middle school student. The prompts were unchanged for each model—the only difference was the model used. The responses of PeerChat were then analyzed on the following four criteria:

- PeerChat's response must be relevant (on-topic)
- PeerChat must struggle (make frequent mistakes)
- PeerChat should respond like a middle schooler (limited academic vocabulary, occasional use of filler words, etc.)
- PeerChat should remember the conversation (consider what has been said so far and respond accordingly)

We developed these criteria to encapsulate the key aspects of a realistic response from a middle school student, since there is limited literature describing how tutees respond in these environments. With this system, we then conducted 16 complete tutoring sessions on the aforementioned problem, giving us enough samples to reasonably quantify each model's success. These 16 tutoring sessions brought a total of 100 back-and-forth interactions with each model, adding to a total of 300 interactions across all models. All 300 interactions were then evaluated as a success or failure on each of the 4 criteria. Subsequent tutoring sessions did not use any data from past sessions.

Results

Figure 2 shows the quality of each model on each criteria by displaying the percentage of utterances where the model had a success on the given criteria.

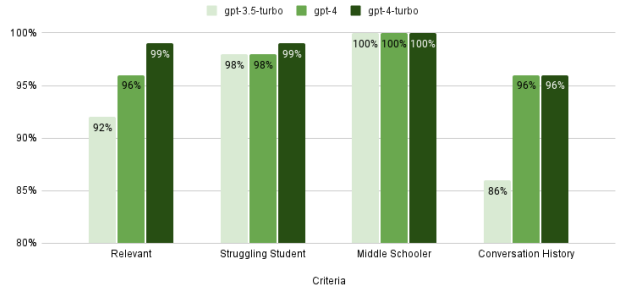

Fig. 2. Evaluation of each model on the four criteria

It appears that gpt-3.5-turbo already achieves considerable accuracy, and that each successive model increases the quality of the responses as evaluated on the given criteria. This suggests that such a system can be built to satisfactory accuracy using prompt engineering on ChatGPT's models.

Conclusion

This paper presents a novel use of generative AI by simulating a tutee in a peer tutoring session and shows the effectiveness of such a system on three of OpenAI's ChatGPT models. Although the system seems to work sufficiently without the need of fine-tuning, further investigations on improving the model (including models that personalize with each interaction) will be necessary to fully realize the potential of AI tutees. Studies on implementing AI tutees with students and tracking relevant outcomes will be crucial in evaluating the effectiveness of such a system. Lastly, a larger analysis with a wide range of topics will help understand the similarities and differences in the kinds of mistakes made by generative AI models compared to students, which is an important aspect of effectively simulating tutees.

References

Bowman-Perrott, L., Davis, H., Vannest, K., Williams, L., Parker, R., & Greenwood, C. 2013. Academic benefits of peer tutoring: A meta-analytic review of single-case research. School Psychology Review, 42(1), 39-55.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. 2020. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901.

Carnegie Learning. 2010. Cognitive tutor effectiveness. Retrieved from http://www.carnegielearning.com/static/web docs/2010 Cognitive Tutor Effectiveness.pdf

Chase, C.C., Chin, D.B., Oppezzo, M.A. et al. 2009. Teachable Agents and the Protégé Effect: Increasing the Effort Towards Learning. J Sci Educ Technol 18, 334–352. https://doi.org/10.1007/s10956-009-9180-4

Kobayashi, K. 2019. Learning by preparing-to-teach and teaching: A meta-analysis. Japanese Psychological Research, 61(3), 192-203.

Lepper, M. R., & Woolverton, M. 2002. The wisdom of practice: Lessons learned from the study of highly effective tutors. In J. Aronson (Ed.), Improving academic achievement: Impact of psychological factors on education (pp. 135-158). Academic Press.

Leung, K. C. 2015. Preliminary empirical model of crucial determinants of best practice for peer tutoring on academic achievement. Journal of Educational Psychology, 107(2), 558–579.

Markel, J. M., Opferman, S. G., Landay, J. A., & Piech, C. 2023. GPTeach: Interactive TA Training with GPT Based Students.

Matsuda, N., Cohen, W.W. & Koedinger, K.R. 2015. Teaching the Teacher: Tutoring SimStudent Leads to More Effective Cognitive Tutor Authoring. Int J Artif Intell Educ 25, 1–34. https://doi.org/10.1007/s40593-014-0020-1

OpenAI. 2023. "ChatGPT," https://openai.com/blog/chatgpt.

Polverini, G; Gregorcic, B. 2023. Performance of a Large Multimodal Model-based chatbot on the Test of Understanding Graphs in Kinematics. arXiv:2311.06946

Tate, T. P., Doroudi, S., Ritchie, D., Xu, Y., & uci, m. w. 2023. Educational Research and AI-Generated Writing: Confronting the Coming Tsunami. https://doi.org/10.35542/osf.io/4mec3

Taylor, R. 1980. The computer in the school: Tutor, tool, tutee.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35:24824–24837.