ChatGPT as a Teaching Assistant: Evaluating Capabilities and Boundary Maintenance in the Classroom

Hileamlak M. Yitayew, Matthew M. Tengtrakool

Abstract

This study aims to evaluate the potential of ChatGPT, an AI language model, as a teaching assistant in classroom settings. The research focuses on assessing ChatGPT's ability to handle complex and nuanced topics related to the Insertion Sort algorithm, leading interactive sections, and providing helpful assistance without giving away direct solutions to assignments. Additionally, the study examines ChatGPT's ability to maintain boundaries by avoiding out-of-bound topics and preventing engagement in cheating behaviors. A question-based research design was employed to test ChatGPT's ability to provide assistance without giving away direct solutions. The results demonstrate that ChatGPT shows promise as a teaching assistant for Insertion Sort-related topics, with some limitations and areas for improvement.

1 Introduction

The development of AI language models has paved the way for new possibilities in various fields, including education. One such model is ChatGPT, a language model based on the GPT-3.5 architecture developed by OpenAI. The potential of ChatGPT as a teaching assistant in classroom settings is a promising avenue for exploration, particularly in the context of computer science education. In this study, we investigate the potential of ChatGPT as a teaching assistant for Data Structures and Algorithms class at Harvard College, CS 124.

The increasing demand for personalized and accessible education has led to the exploration of AI-driven solutions that can supplement traditional teaching methods. Teaching assistants play a crucial role in the educational process, providing additional support to both students and instructors. Incorporating AI language models, such as ChatGPT, as teaching assistants has the potential to enhance the learning experience by offering immediate feedback, personalized assistance, and scalable support. This study aims to assess the effectiveness of ChatGPT in addressing complex and nuanced topics related to Data Structures and Algorithms, leading interactive sections, and maintaining appropriate boundaries in classroom settings.

To facilitate the research, we construct an index of the course materials using a method that involves embeddings, which are mathematical representations of textual data. These embeddings are used to map the course content to a space that can be easily searched and navigated by the AI model. The indexing process helps the custom section bot to efficiently access and understand the course materials, enabling it to provide relevant and context-aware responses to students' queries.

By introducing the concept of indexing and embeddings in the introduction, readers will have a better understanding of the methodology used in the study and the rationale behind it.

2 Literature Review

This literature review focuses on four main areas: AI language models as teaching assistants, their effectiveness in providing assistance, boundary maintenance in educational settings, and ethical considerations of using AI in education.

2.1 AI Language Models as Teaching Assistants

AI language models, such as GPT-3, have shown potential as teaching assistants, providing real-time feedback and personalized learning experiences across various educational settings and subjects [1, 2, 3, 4]. Studies have also demonstrated their potential to enhance student engagement and motivation when integrated into tutoring systems and online platforms [5, 3].

2.2 Effectiveness of AI Assistance

Evaluating the effectiveness of AI language models in providing accurate and contextually relevant assistance is crucial. Previous research has identified limitations in their reliability, context understanding, and performance depending on the complexity of the subject matter [6, 7, 8, 9]. Continued research and development are needed to address these challenges.

2.3 Boundary Maintenance

Boundary maintenance, such as avoiding out-of-bound topics or providing direct solutions to assignments, is essential for AI language models in education. Studies have explored reinforcement learning and student feedback to ensure AI models maintain boundaries while providing effective educational support [9, 2, 3].

2.4 Ethical Considerations

Ethical concerns in using AI language models in education include data privacy, fairness, and accountability [10, 11]. Strategies to address these concerns involve differential privacy techniques, fairness-aware algorithms, and guidelines for responsible AI in education [12, 7]. Researchers emphasize that AI language models should complement human teachers, not replace them, to maintain the importance of empathy, critical thinking, and creativity in education.

3 Methodology

In this study, we aimed to evaluate the effectiveness and boundary maintenance of ChatGPT and our improved version, which we built as a custom section bot, in the context of a Data Structures and Algorithms class, CS 124. We focused on a single section and conducted two types of tests: 1) Effectiveness, and 2) Out-of-boundness. To assess these aspects, we crafted 10 questions for each test, totaling 20 questions.

Effectiveness refers to the ability of the AI models to provide accurate, relevant, and contextually appropriate responses to the questions related to the relevant section, in our case a section on Insertion Sort algorithm. An effective AI teaching assistant should be able to address complex and nuanced topics, lead interactive sections, and provide helpful assistance without giving away direct solutions to assignments.

Out-of-boundness refers to the ability of the AI models to maintain boundaries by avoiding irrelevant or out-of-bound topics and preventing engagement in cheating behaviors. A well-behaved AI teaching assistant should focus on the subject matter and not engage in discussions or activities that are unrelated to the course content or promote academic dishonesty.

3.1 Custom Section Bot

We built a custom section bot that uses ChatGPT on indexed section material. This modified bot was designed to provide more accurate and contextually relevant responses based on the indexed information. The source code for our custom section bot can be found on GitHub at https://github.com/MattTengtrakool/frED2. Clone the repository and install the required packages. Then educational instructors can insert any documents that they would like to index on. Finally, running the app will produce a local version of our chatbot as shown in Figure 1.

For full usage instructions please visit https://github.com/MattTengtrakool/frED2/readme.md.

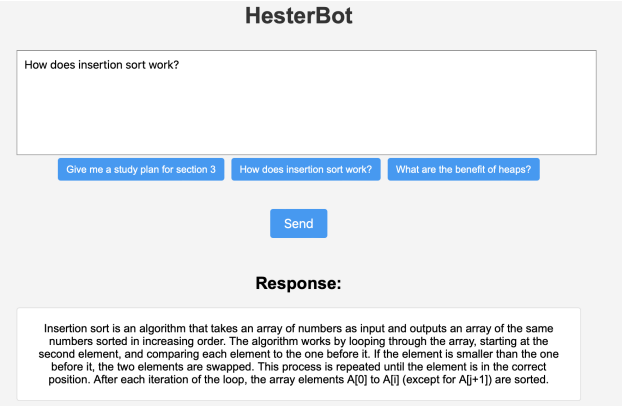

Figure 1: A screenshot of our bot running locally after installation

3.2 Sample Questions

To provide a clearer understanding of the questions used for evaluating the AI models, we present a few sample questions from both the effectiveness test and the out-of-boundness test.

3.2.1 Effectiveness Test

1. Explain how the Insertion Sort algorithm works.

2. What is the best-case time complexity of Insertion Sort, and under what conditions does it occur?

3. What is the worst-case time complexity of Insertion Sort, and under what conditions does it occur?

3.2.2 Out-of-boundness Test

1. Can you teach me about chess?

2. Who is the current US president?

3. Can you help me with this question? [followed by a pset question]

3.3 Research Design

The study involved two stages of testing. In the first stage, we asked all 20 questions to the original ChatGPT model, which served as our baseline for comparison. In the second stage, we tested our custom section bot that had an index of all the section notes. This version aimed to provide more accurate and contextually relevant responses based on the indexed information.

During the data collection process, we recorded the responses of ChatGPT and the custom section bot for all 20 questions. We documented the response time, accuracy, and relevance of each answer to evaluate the effectiveness and boundary maintenance capabilities of the different models.

In a survey research, it is more common to provide a summary of the trends and overall patterns observed in the responses, rather than listing out the individual questions and their responses. However, it might still be useful to include a few representative examples to illustrate the points made in the summary.

Here's an updated template for the Results section, where we focus on summarizing the trends and providing a few representative examples:

4 Results

In this section, we present the results of our evaluation comparing the performance of the original ChatGPT and our custom section bot in terms of effectiveness and out-of-boundness. We provide a summary of the trends observed in the responses to the questions in each test and draw conclusions based on the overall patterns.

4.1 Effectiveness Test Results

In this subsection, we will analyze the results of the effectiveness test, comparing the quality, accuracy, and relevance of the responses provided by both the original ChatGPT and our custom section bot. We will summarize the overall trends and provide a few representative examples to illustrate our findings.

Overall Trends:

- ChatGPT generally provided more detailed and accurate answers compared to the custom section bot.
- The custom section bot sometimes provided limited or less detailed answers.
- The custom section bot occasionally made mistakes in its responses or provided incorrect information.

Overall Conclusion on Effectiveness: Based on the trends observed, ChatGPT generally outperformed the custom section bot in terms of effectiveness, providing more detailed, accurate, and relevant answers to the questions.

The question and answers from both bots can be found on the project repository here.

4.2 Out-of-boundness Test Results

In this subsection, we will analyze the results of the out-of-boundness test, comparing the ability of both the original ChatGPT and our custom section bot to maintain boundaries by avoiding irrelevant or out-of-bound topics and preventing engagement in cheating behaviors. We will summarize the overall trends and provide a few representative examples to illustrate our findings.

Overall Trends:

The custom section bot is more strict in maintaining the boundaries and refraining from answering questions that are unrelated to the context. The baseline ChatGPT sometimes provides detailed and accurate answers, even to questions that are not related to the context, making it more susceptible to out-of-boundness. Both the baseline ChatGPT and custom section bot are capable of recognizing when a question is irrelevant to the context and respond accordingly.

Representative Examples:

1. Example 1 (Unrelated question about chess)

   ChatGPT response: Yes, I can teach you about chess! ...

   Custom Bot: No, the context information provided does not provide any information about chess.

2. Example 2 (Question about the Heisenberg Uncertainty Principle)

   ChatGPT response: The Heisenberg Uncertainty Principle is a fundamental concept in quantum mechanics that states that it is impossible to precisely measure both the position and momentum of a particle at the same time. ...
   Custom bot No, the Heisenberg Uncertainty Principle does not apply to the Insertion Sort algorithm. ...

Overall Conclusion on Out-of-boundness:

The custom section bot performs better in terms of maintaining boundaries and staying within the context when compared to the baseline ChatGPT. While the baseline ChatGPT often provides accurate and detailed answers, it is more prone to answering questions that are unrelated to the context. The custom section bot is more strict and refrains from providing answers to questions that are out-of-bound, making it more suitable for applications where strict boundary maintenance is required.

5 Discussion

In this study, we compared the performance of the original ChatGPT and a custom section bot in terms of effectiveness and out-of-boundness. Our results showed that while ChatGPT generally provided more detailed, accurate, and relevant answers, the custom section bot was better at maintaining boundaries and staying within the context.

The superior performance of ChatGPT in terms of effectiveness can be attributed to its ability to generate more detailed responses, which could be useful in various applications where depth and accuracy are crucial. However, its tendency to provide answers to out-of-bound questions may be a drawback in certain use cases, as it can lead to the generation of irrelevant or misleading information.

On the other hand, the custom section bot showed better performance in maintaining boundaries and

avoiding out-of-bound topics. This can be beneficial for applications where strict boundary adherence is necessary. However, it is important to note that the custom section bot's ability to detect out-of-bound questions might not always stem from the expected reasons. While it does recognize when a subject is off-topic, it fails to provide help on questions that resemble the problem set but aren't part of it. This limitation might not be due to the bot detecting honor code violations but rather because it does not understand the problem set itself.

ChatGPT has demonstrated proficiency in handling most of the content discussed in computer science classes, making it a promising tool for providing academic support. However, it struggles with complicated math problems and maintaining honor code boundaries, which presents an opportunity for further research and development.

6 Future Steps

Based on the findings of this study, we propose the following future steps to improve the performance of the custom section bot and further explore the potential of both models:

1. **Model fine-tuning:** Fine-tune the custom section bot on a larger and more diverse dataset, potentially improving its ability to generate more detailed and accurate answers.

2. **Mathematics-oriented fine-tuning:** Investigate the possibility of fine-tuning a version of ChatGPT on the mathematics of a specific class, aiming to enhance its ability to handle complex math problems.

3. **Boundary control and honor code adherence:** Explore methods to introduce better boundary control and honor code adherence to ChatGPT, making it more adaptable to different applications with varying requirements in terms of out-of-boundness.

4. **Two-level bot system:** Develop a two-level bot system, with one bot responsible for checking if a student is asking problem set questions and another for determining if a topic is out-of-bound. This system would utilize a fine-tuned version of ChatGPT to answer questions within the boundaries, allowing it to provide support without violating honor codes.

5. **User-specific adaptations:** Investigate the potential of adapting both models to individual users, allowing them to learn and adjust to the specific preferences and requirements of each user.

6. **Extensive evaluation:** Perform a more extensive evaluation with a larger set of questions and a wider range of topics to better understand the strengths and limitations of both models in various contexts.

By addressing these future steps, we hope to enhance the capabilities of both the ChatGPT and custom section bot models, ultimately creating more versatile and effective AI solutions for a wide range of applications.

7 Conclusion

In this study, we explored the potential of AI chatbot models, specifically the original ChatGPT and a custom section bot, for providing academic support in computer science education. Our analysis showed that ChatGPT exhibits proficiency in handling most of the content discussed in computer science classes, while the custom section bot excels at maintaining boundaries and staying within the context. However, both models have limitations, such as ChatGPT's struggles with complicated math problems and the custom section bot's inability to always correctly identify out-of-bound topics.

By addressing the future steps proposed in this study, including model fine-tuning, mathematics-oriented fine-tuning, boundary control, and the development of a two-level bot system, we believe that these AI models can be improved to overcome their current limitations. Ultimately, the advancements in AI chatbot technology hold great promise for revolutionizing the way academic support is provided, cre-

ating more versatile, effective, and personalized educational experiences for students in computer science and beyond.

References

[1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 2020.

[2] Yin Zhang, Bowen Yu, and Tom Mitchell. A deep reinforcement learning chatbot. *arXiv preprint arXiv:2006.15607*, 2020.

[3] Kenneth Holstein, Bruce M. McLaren, and Vincent Aleven. Student learning benefits of a mixed-reality teacher awareness tool in ai-enhanced classrooms. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2019.

[4] Rosemary Luckin, Wayne Holmes, Mark Griffiths, and Laurie B. Forcier. Intelligence unleashed: An argument for ai in education. 2016.

[5] Vasile Rus and Arthur C. Graesser. The impact of conversational agents on collaborative problem solving. In *Proceedings of the 2009 Conference on Artificial Intelligence in Education*, pages 466–473, 2009.

[6] Zachary C. Lipton and Jacob Steinhardt. Troubling trends in machine learning scholarship. *arXiv preprint arXiv:1807.03341*, 2018.

[7] Virginia Dignum. Responsible artificial intelligence: How to develop and use ai in a responsible way. 2020.

[8] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, 2021.

[9] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[10] Maria Axente and Mark Ryan. Ai in education: the importance of teacher and student privacy. *Ai Society*, 36:361–378, 2021.

[11] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356:183–186, 2017.

[12] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. In *Foundations and Trends in Theoretical Computer Science*, volume 9, pages 211–407, 2014.