

Tracking the Evolution of Student Interactions With an LLM-Powered Tutor

Kevin Gold, Shuang Geng
Boston University, USA
klgold@bu.edu, sgeng@bu.edu

ABSTRACT: Student usage of an LLM-powered tutor to get homework help was tracked over the course of a semester in a university-level introductory data science course. For each homework assignment, the GPT-4 powered tutor was given the text of the homework problems and solutions in advance but was instructed to never reveal solutions directly, instead guiding the student to the correct answer through leading questions. Despite the free availability of ChatGPT, the majority of the class used the system. Anonymous logs were coded on seventeen dimensions of interaction. Evidence indicates that the students found the bot nearly as helpful as the human teaching assistants (TAs), and the bot was utilized more than the TAs' office hours. Some patterns of misuse, such as using the bot for convenient code checking, increased over the course of the semester.

Keywords: AI Tutoring, Large Language Models (LLMs), ChatGPT, Educational Technology, Learning Analytics, Qualitative Content Analysis, Higher Education, Programming Education.

1. INTRODUCTION

A looming challenge that large language models (LLMs) pose to the traditional homework model is that students could simply ask an LLM to do assignments for them. Yet, LLMs have the potential to enable a degree of personal attention that would have been previously impossible at scale (Wu et al., 2023). This work explores the latter model of homework, where an AI tutor provides a steady supply of helpful hints without jumping too soon to the answer. While early work investigating the use of ChatGPT in education has rendered mixed results of varied performance across subject domains (Lo, 2023), we focus on evaluating the performance of using a GPT-4 API as an AI tutor in an introductory data science course at Boston University in Fall 2023. This tutor is given the homework assignment with solutions as part of its context, enabling convenient queries such as "How do I approach 1b," but is instructed to use the Socratic method to dole out guidance more frugally than ChatGPT. Three research questions about student's interactions with the tutor will be addressed: **1) How helpful can LLMs be in tutoring college students? 2) What are the typical bot failures and their trends over time? 3) What are the typical student behaviors and their trends over time?**

2. DATA & METHODOLOGY

The AI tutor provided a graphical user interface (GUI) for students to select an assignment via a radio button and enter a query. The student's query was augmented with a prefix: "For this query, answer with a single question that you haven't asked before that is meant to lead someone in the right direction, without directly answering the relevant homework question - unless the problem is solved completely, in which case, quit." The query was also augmented with the system information, "You

are a helpful teaching assistant in a data science course. Your primary goal is to help the students learn. This is the homework the student was talking about: [homework text & solution]".

Two data sources are evaluated. The first one is a set of chatlogs of student-AI interactions. 802 sessions were recorded in total on eight homework assignments in a class of 127 students. We used a combination of deductive and inductive coding approaches to construct a set of initial interaction classifications based on chatlogs of the first two assignments. Our coding unit is each *problem* rather than *session* as students can ask about multiple homework problems in one session. Our final codebook includes 17 dimensions under four parent categories: A) Helpfulness of Advice; B) Bot Failures: Leak Correct Answer, Clear Error in Answer, Provide Irrelevant Answer, Bot Demands Extra Work, Fail to Point Back to Course Materials; C) Student Misuses: Select Wrong HW in GUI, Spam for Hints, Unclear Prompt, Search for Exploits; and D) Student Question Types: Debug Request, Review Code, Improve Style, Clarify Concept, Ask for Example, Recommend Resource, and General Hint Request. All dimensions are Boolean variables except A and B1, which are ordinal. We sampled 10 sessions from each of HW2 to HW7 for coding, with a total of 60 out of 802 sessions and 152 problems. (The easy HW1 mostly saw frivolous student behaviors, and HW8 was optional, so we excluded both.)

The second data source is a pair of surveys on students' experience with the AI tutor. 50 midterm survey forms and 65 end-semester survey forms were collected from 127 students.

3. RESULTS

1) How helpful can LLMs be in tutoring college students? Evidence indicates that most of the bot's answers are deemed very or moderately helpful by both the students (in surveys) and the independent coders with sufficient programming backgrounds (using chatlogs). Table 1 shows the percentage breakdowns before and after the midterm and across the data sources. In the end survey, students rank the AI tutor's helpfulness as equal to that of human tutors.

Table 1: Helpfulness of the AI Tutor

	Before Midterm		After Midterm	
	Chatlogs	Midterm Survey	Chatlogs	End Survey
Helpfulness (Very/Moderate/Not)	86% / 9% / 5%	29% / 62% / 9%	84% / 8% / 8%	42% / 49% / 9%

While students find the AI tutor helpful in learning, they also utilize it more than the TA office hours. 69% of the respondents report that they have used the bot at least once, compared to only 46% who have visited TA office hours. Perceived ease of use, usefulness, and convenience have been identified as main factors of AI tutors' preferability (Malik et al., 2021).

2) What are the typical bot failures and their trends over time? Overall, we see a decreasing trend of bot failures in later assignments. Fisher's exact tests were performed across the five failure categories to discern significant trends, and statistically significant results were observed on B1-Leak Correct Answer ($p\text{-value}=1.60\text{e-}04$). Figure 1 (left) shows boxplots of the 4 levels of answer leaking across homework (4 levels: 0=no leak to 3=verbatim answer). The average level changes from 1.20 (near 1=mild leak) in HW2 to 0.34 in HW7 (almost no leak), showing that as problems became more multipart, the bot was less likely to leak the full problem.

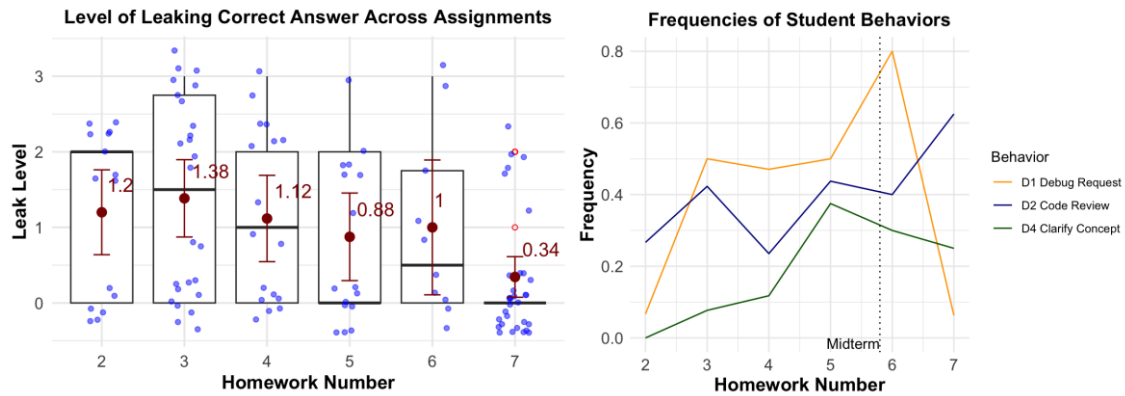


Figure 1: Evolution of interactions. Left: Leak Answer Trend. Right: Freq. of Student Behaviors

3) What are the typical student behaviors and their trends over time? Fisher's exact tests were performed across student misuse cases and question types to discern significant trends before and after the midterm. Statistically significant results were observed on D1-Debug Request ($p=4.32e-05$), D2-Review Code ($p=4.19e-02$), and D4-Clarify Concept ($p=3.47e-02$). The frequency of debug requests (asking the bot what is wrong with code in progress) went down from 0.46 to 0.08 (pre- to post-midterm), while that of code reviews (asking whether complete code is correct) went up from 0.36 to 0.58. One possible explanation is that the return of the graded in-class midterms generally convinced the class that they should not be overly reliant on the bot to debug programs (see Figure 1 right, showing frequencies by homework of the behaviors with statistically significant changes).

4. CONCLUSIONS

The evidence generally points to the AI tutor's advice being perceived as helpful by both students and coders evaluating the interactions. However, the evidence also suggests that misuse can rise over time as students learn to use the system in unintended ways, such as using it as a convenient answer checker or a "debugger" that is powerful enough to write the program one debugging step at a time (though in-class midterms serve as a deterrent to this behavior). It has been suggested that debugging is itself a useful activity for exercising critical thinking about a program and building the students' mental model of the program, the "notional machine" (Lowe, 2019). Further research will look at how to intercept or flag problematic interactions as they happen.

REFERENCES

- Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences*, 13(4), 410. <https://doi.org/10.3390/educsci13040410>
- Lowe, T. (2019). Debugging: The key to unlocking the mind of a novice programmer? *2019 IEEE Frontiers in Education Conference (FIE)*, 1-9. <https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9028699>
- Malik, R., Shrama, A., Trivedi, S., & Mishra, R. (2021). Adoption of chatbots for learning among university students: Role of perceived convenience and enhanced performance. *International Journal of Emerging Technologies in Learning (iJET)*, 16(18), 200-212.
- Wu, T.-T., Lee, H.-Y., Li, P.-H., Huang, C.-N., & Huang, Y.-M. (2024). Promoting Self-Regulation Progress and Knowledge Construction in Blended Learning via ChatGPT-Based Learning Aid. *Journal of Educational Computing Research*, 61(8), 3-31. <https://doi.org/10.1177/07356331231191125>