Could ChatGPT be an Engineer? Evaluating Higher Education Vulnerability to AI Assistants

Anonymous submission

Abstract

With the release of ChatGPT, the incredible potential of Large Language Models (LLMs) to perform a wide array of tasks has been seared into the public mind, inviting both excitement and concern about the significant changes caused by widespread LLM usage. This paper investigates how grounded these concerns are by investigating how much university students can leverage these models to answer STEM courses' questions and problems. We examine the abilities of GPT-3.5 and GPT-4 in a bilingual college-level education setting by having them answer questions from $\sim$100 of our university's courses across a variety of subjects. We employ state-of-the-art prompting strategies and analyze the obtained results across different axes. Using both automatic and human grading, through our university's teaching staff, we find that GPT-4 consistently outperforms GPT-3.5, with the latter being freely available to the public and able to pass 34% of the courses it was tested on. We observe that the models' performance is affected not only by the prompting strategy but also by course topic and language. While they perform better in English, this is not necessarily an impediment to their performance in other languages. We also find introductory and general courses to be more susceptible to LLMs, thought they struggle with uncommon question formats and questions that require multi-step reasoning. We conclude all courses have some level of vulnerability to LLMs. On the other end of applying LLMs to educational domains, we analyze GPT-4's potential as an automatic grader. We find it insufficient compared to human graders, in part because of its tendency to avoid marking answers as either definitively correct or incorrect. Finally, we provide a set of implications and takeaways for educators on to make their course material less susceptible to the challenges posed by LLMs' usage.

Introduction

ChatGPT was released in November 2022 to broad adoption and fanfare, reaching 100M users in its first month of use, and remaining in the public discourse to this day. Its release has prompted a continuing discussion of societal transformations likely to be induced by AI over the next years and decades. One core area of this discussion has been the domain of education, assessing the risk of these systems being used as cheating tools that would allow students to succeed at assessments without learning the relevant skills.

Despite this concern, there has yet to be a comprehensive empirical study of the potential impact of LLMs on education institutions' assessment methods. While a few studies have explored the interesting subtask of how well models perform on different course topics (Hendrycks et al. 2021; Huang et al. 2023; Wang et al. 2023b; Zhong et al. 2023; Arora, Singh, and Mausam 2023), and aggregated relevant question sets for this purpose, none have extrapolated these findings to education systems and degree programs.

Contributions (1) We compile a broad (94 subjects) challenging dataset of English and French STEM questions, where the majority of questions require multi-hop reasoning and cover all levels of higher education. (2) We conduct a comprehensive study of the performance of prompting strategies on our dataset.(3) We analyze the performance of the different prompting strategies on many axes, including model, question topic, and language, among others. We find topic and language to be the most impactful ones after prompting strategy. (4) We evaluate GPT-4's potential as a grader by comparing its performance against the actual grading of these courses' teaching staff. (5) We discuss the main implications and takeaways of our findings for educators, and open-source our code for model prompting, analysis, and grading.

Experimental Setup

Dataset We compile a novel dataset comprising data sourced from 94 different classes at our university. Our dataset includes data from both on-campus classes as well as online classes covering a wide range of science, technology, engineering, and mathematics domains.. After data preprocessing and filtering, this results in a bank of 6442 English and French questions — 4039 Multiple Choice Questions (MCQs) and 2403 open-answer questions.

Prompting Strategies We consider three main categories of prompting strategies: *direct prompting*, where the model is asked to directly provide an answer to the question (**Zero-shot**, **One-shot** (Brown et al. 2020), and **Expert Prompting** (Xu et al. 2023)); *rationalized prompting*, where the model is encouraged to first reason about the problem before arriving at an answer (**Chain-of-Thought** (CoT; Wei et al. 2023), and **Tree-of-Thought Prompting** (Yao et al. 2023)); *reflective prompting*, where the model is asked to reflect on a previously generated response before deciding on the final answer (**Self-Reflect** (Wang et al. 2023a; Madaan et al. 2023),

and **Metacognitive Prompting** (Wang and Zhao 2023)).

Evaluation Setup We employ different evaluation strategies depending on the question type. MCQs can be automatically graded by extracting the answer index(es). To grade open-answer questions, we perform direct answer grading using GPT-4's ability to understand the question, the solution, and the provided answer. In parallel, we also have human experts grade the model's answers so we can establish a confidence level for the model-assigned grades.

Analysis

Overview Using our university's passing threshold of 67% or higher, GPT-4 would pass 93 out of 142 our courses, while GPT-3.5 would pass only 49. We find model performance is always sensitive to the question language and topic, and usually affected by question difficulty and generality.

Question Difficulty We find a progressive drop in performance as difficulty and cognitive load increase. Both GPT-4 and GPT-3.5 seem to have a fairly low correlation between their performance and the average student performance. Surprisingly, we observe that GPT-3.5 appears to outperform GPT-4 in complementing a student's performance, as it achieves higher performance in questions where the average student performance was low.

Course Generality Model performance has some degree of correlation with courses designed for wider audiences. One possible reason for this is that large courses, designed for hundreds of students, are either introductory in nature or popular enough that there are also more resources to learn from — and as such, they are particularly vulnerable to model usage. This can be counteracted when less specialized courses have a bigger presence of math and other specialized domains whose question and answer formats hinder model performance (e.g., generating compilable LaTeX code).

Question Language GPT-4 is better on English MCQs than French ones. Assuming questions are not translated into English, we find three important results. First, for complex instructions, the language of the instruction greatly impacts the language of the generation. Second, providing the instruction in the same language as the question leads to higher performance for both models compared to when the question is in a different language. Finally, for questions in French, the models generally achieve higher performance scores when they answer in the same language.

Course Topic We find a significant difference between MCQs and open-answer performance — however, their different grading strategies must be taken into account. MCQs are graded through an exact match, giving us a more precise grade than open-answer questions, graded by GPT-4. GPT-3.5 is consistently outperformed by GPT-4. Despite the significant variation, *Computer Software* and *Computer Systems* appear to be among the best-performing topics, while *Mathematics*, *Applied Computer Science*, and *Linguistics* are among the topics GPT-4 is weakest at. While the models seem to perform best for areas for which there are more resources available on the internet and for straightforward generation formats (text or code), all courses have some degree of vulnerability, regardless of their subject.

Grading experts' remarks Graders impressions of the models range from acceptable to very poor, often depending on the type of question. There was wide consensus that the generations were decent for simple questions, but not for any questions that required logical reasoning or analysis. In those cases, they found the models wrote long answers deprived of any actual information, made circular arguments or used implications that were not valid.

Impact of Prompt Strategy GPT-4 outperforms GPT-3.5 across all prompting strategies. When answering MCQs, four-shot CoT emerges as GPT-4's best-performing strategy, while zero-shot is the worst one. Curiously, the same ranking does not transfer to open-answer questions, where Self-Reflect emerges as the best strategy, followed by Expert Prompting. Zero-shot prompting remains the worst prompting strategy. Based on a survey of reports submitted by Master students in the context of a class in Modern Natural Language Processing, we found students to be most likely to use Zero-shot, Expert, Zero-shot COT prompting strategies, as these are the most intuitive strategies, as well as the ones that require the least amount of preparation work.

GPT-4-based Grading We find that while the average and aggregate statistics statistics might suggest GPT-4 can be used as a grader, a deeper look reveals its inability to mimic the grading behavior and distributions of actual human graders. We conclude that, in its current state, GPT-4 should never be employed as an automatic grader.

Conclusion and Takeaways

We found courses which are more introductory or designed for bigger audiences to be more vulnerable to LLMs. This might also be due to the higher number of resources available on these topics. Conversely, we found that courses relying on non-textual and non-code formats to be harder for models to answer correctly (e.g., math questions relying on LaTeX). Nevertheless, all courses and topics had some level of vulnerability, regardless of how hard the subject matter was deemed by experts. Models perform better in English questions, but a variety of strategies (e.g., translating the questions in English) can make courses in other languages equally susceptible to model usage. While our expert graders from the teaching staff found CoT to perform worse than Metacognitive prompting, students are most likely to use Zero-shot, Zero-shot CoT and Expert prompting. The best strategies for preventing LLMs exploitation by students lie at both ends of the spectrum: either closed-book exams or very open assignments, such as analysis reports and research projects. Finally, though GPT-4 can appear to be a reliable grader when looking only at aggregate statistics, it has a very low correlation with actual human expert grading, and should not be employed as an automatic grader.

References

Arora, D.; Singh, H. G.; and Mausam. 2023. Have LLMs Advanced Enough? A Challenging Problem Solving Benchmark For Large Language Models. arXiv:2305.15074.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; 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.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 1877–1901. Curran Associates, Inc.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2021. Measuring Massive Multitask Language Understanding. arXiv:2009.03300.

Huang, Y.; Bai, Y.; Zhu, Z.; Zhang, J.; Zhang, J.; Su, T.; Liu, J.; Lv, C.; Zhang, Y.; Lei, J.; Fu, Y.; Sun, M.; and He, J. 2023. C-Eval: A Multi-Level Multi-Discipline Chinese Evaluation Suite for Foundation Models. arXiv:2305.08322.

Madaan, A.; Tandon, N.; Gupta, P.; Hallinan, S.; Gao, L.; Wiegreffe, S.; Alon, U.; Dziri, N.; Prabhumoye, S.; Yang, Y.; Gupta, S.; Majumder, B. P.; Hermann, K.; Welleck, S.; Yazdanbakhsh, A.; and Clark, P. 2023. Self-Refine: Iterative Refinement with Self-Feedback. *arXiv:2303.17651*.

Wang, R.; Wang, H.; Mi, F.; Chen, Y.; Xu, R.; and Wong, K.-F. 2023a. Self-Critique Prompting with Large Language Models for Inductive Instructions. arXiv:2305.13733.

Wang, X.; Hu, Z.; Lu, P.; Zhu, Y.; Zhang, J.; Subramaniam, S.; Loomba, A. R.; Zhang, S.; Sun, Y.; and Wang, W. 2023b. SciBench: Evaluating College-Level Scientific Problem-Solving Abilities of Large Language Models. arXiv:2307.10635.

Wang, Y.; and Zhao, Y. 2023. Metacognitive Prompting Improves Understanding in Large Language Models. arXiv:2308.05342.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; and Zhou, D. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903.

Xu, B.; Yang, A.; Lin, J.; Wang, Q.; Zhou, C.; Zhang, Y.; and Mao, Z. 2023. ExpertPrompting: Instructing Large Language Models to be Distinguished Experts. arXiv:2305.14688.

Yao, S.; Yu, D.; Zhao, J.; Shafran, I.; Griffiths, T. L.; Cao, Y.; and Narasimhan, K. 2023. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. arXiv:2305.10601.

Zhong, W.; Cui, R.; Guo, Y.; Liang, Y.; Lu, S.; Wang, Y.; Saied, A.; Chen, W.; and Duan, N. 2023. AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models. arXiv:2304.06364.