Can You Trust a Generative AI Teacher? Verifying the Imperfections of Generative AI for Educational Purposes Using a Zero-Shot Level Classification Problem

Anonymous submission

Abstract

Recently, with the advent of GPT-4 and GPT Store, various educational LLMs have been utilized. However, from an educational and social perspective, we felt the need to carefully examine this phenomenon. One of the most basic factors in education is to identify the level of learners, and for this purpose, teachers should be able to determine the hierarchy and level of knowledge. Our research team used the ARC-E and ARC-C data to define a science question and answer level classification problem, and through experiments, we found that the current LLM still has limitations in clearly distinguishing the hierarchy and level of knowledge. From an educational perspective, this result strongly suggests that using LLM for educational purposes may make it difficult to provide appropriate education at the learner's level, which may undermine the credibility of education.

Introduction

With the rapid development of generative AI in society, it is rapidly being adopted in various fields. Education is one of the most prominent areas of adoption. Students are now utilizing AI in a variety of ways, including assignments and learning. However, our research team raises a question. In the real world, teachers are able to identify students' inclinations and levels, and the data underlying these judgments are made and refined through student utterances, such as questions. We designed a binary classification problem using the ARC data set and evaluated its performance through zero-shot learning on GPT-4 to verify whether the current generative AI can perform this process. As a result, depending on the amount of various data sets, GPT-4 did not perform as expected, which means that even GPT-4, which is a representative of generative AI, is good at solving problems but has difficulty classifying the level of the problem. This suggests that it may be difficult for students to learn by level based on their utterances, which further suggests that pedagogically, the use of generative AI may be less effective than human teachers, which raises questions about the educational adoption of generative AI in society.

Method

Designing Experiments

This study aims to evaluate the zero-shot classification ability of large language models (LLMs) in the context of classifying educational questions, specifically using GPT-3.5 and GPT-4. Two datasets, ARC-Easy and ARC-Challenge, representing different levels of question complexity, were used in this experiment. We hypothesized that GPT-3.5 and GPT-4 would be able to effectively distinguish between these two categories without explicit prior training on these specific datasets.

Prepare Data

Two datasets were used in this experiment: ARC-Easy and ARC-Challenge. Each dataset consists of questions from each category. The ARC-Easy dataset contains relatively simple questions (labeled "1"), while the ARC-Challenge dataset contains more complex and challenging questions (labeled "2"). For each dataset, we extracted the first 1000 questions.

Each question in the ARC-Easy dataset was assigned a label of '1' and each question in the ARC-Challenge dataset was assigned a label of '2'. These labels represent the level of complexity of the question and serve as a basis for evaluating the classification accuracy of the model.

Measure Execution and Performance

For this experiment, we used a GPT-3.5, GPT-4 model via the OpenAI API. The model was tasked with classifying each question into a '1' (ARC-Easy) or '2' (ARC-Challenge) category based on its content. Custom prompts for each question were designed to guide the model in its classification.

The experiment was performed with a randomized combination of questions from both datasets. For each question, the model's output was compared to the assigned label to determine accuracy. The experiment was performed repeatedly for each question, with accuracy recorded every 100 questions to observe trends in the model's performance.

The performance of the model was evaluated based on its accuracy in correctly classifying questions into each category. Accuracy was calculated as the number of correct answer predictions divided by the total number of questions processed in each bin.

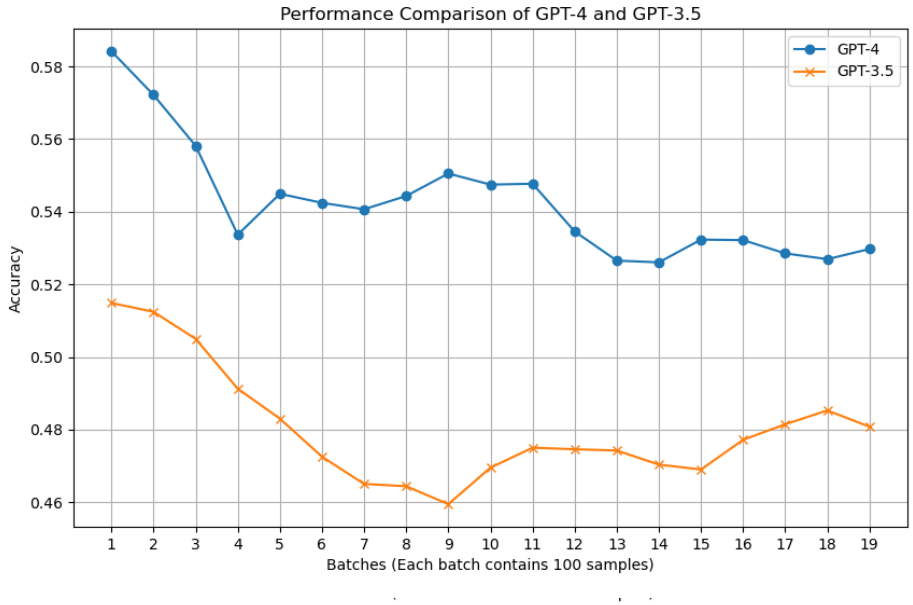

Figure 1: The figure illustrates the classification accuracy of two generative language models, GPT-4 and GPT-3.5, across nineteen batches, each comprising 100 samples from a mixed set of educational questions. GPT-4 maintains a higher accuracy throughout, peaking at the initial batch, while GPT-3.5 shows a decline in performance, with a slight increase in the latter batches.

Experimental Results

Overview

This experiment aimed to evaluate the zero-shot classification ability of GPT-4 and GPT-3.5 on a set of training problems categorized into two difficulty levels. The performance of each model was evaluated based on its accuracy in correctly classifying these problems.

Analyze Data

The study was performed on 2,000 questions, divided equally between the ARC-Easy and ARC-Challenge datasets. The accuracy of the GPT-4 and GPT-3.5 models was calculated at 100-question intervals, yielding 19 data points for each model.

Results

The analysis showed a distinct performance pattern for each model, with the following results
GPT-4 performance: GPT-4 had consistently high accuracy across all batches, peaking at 58.4 in the first batch and then dropping slightly before stabilizing. The accuracy of this model fluctuated gently throughout the experiment, with an average accuracy of around 53.2.
GPT-3.5 performance: GPT-3.5 had lower accuracy compared to GPT-4, starting with 51.5 in the first batch and generally declining in later batches. The performance of this model improved slightly in later batches, but the average accuracy was still lower than GPT-4 at around 47.6.

Conclusions

Although ours was a relatively simple experiment, we believe that the results can suggest a number of things. First of all, we can see that current generative AI has difficulty identifying the level of the problem. Of course, the data used in our study was about 2000 items, and it is characterized by a relatively small amount of data, which is different from the typical supervised learning process. We agree that this can lead to significant differences in the learning and performance of the model. In such a realistic situation, each student's utterances or questions are different, and such data is relatively difficult to obtain and label. Therefore, I think our experiment is a good representation of this situation in light of the realistic situation. And based on the results, we can see that generative AI will inevitably have difficulty identifying the level of the student's utterances. If these limitations of A.I. are not overcome and A.I. is incorporated into education, students will experience education that is not appropriate for their level, which may cause a gap between the real (offline) learning process and the learning process through A.I., which may cause cognitive confusion for students. Such indiscriminate use of A.I. in education is likely to result in social costs due to the decline in the quality of education.
We propose to build on this research and conduct further research to ensure that student utterances, questions, and general hierarchies of textbook knowledge are well learned by generative AI and utilized in student conversations and instruction.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Baird, M. D.; Pane, J. F.; Steiner, E. D.; Hamilton, L. S.; and Pane, J. D. 2017. How does personalized learning affect student achievement?

Boratko, M.; Padigela, H.; Mikkilineni, D.; Yuvraj, P.; Das, R.; McCallum, A.; Chang, M.; Fokoue-Nkoutche, A.; Kapanipathi, P.; Mattei, N.; et al. 2018. A systematic classification of knowledge, reasoning, and context within the ARC dataset. *arXiv preprint arXiv:1806.00358*.

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.

Plaza-del Arco, F. M.; Nozza, D.; and Hovy, D. 2023. Leveraging Label Variation in Large Language Models for Zero-Shot Text Classification. *arXiv preprint arXiv:2307.12973*.

Puri, R.; and Catanzaro, B. 2019. Zero-shot text classification with generative language models. *arXiv preprint arXiv:1912.10165*.

Sridhar, P.; Doyle, A.; Agarwal, A.; Bogart, C.; Savelka, J.; and Sakr, M. 2023. Harnessing llms in curricular design: Using gpt-4 to support authoring of learning objectives. *arXiv preprint arXiv:2306.17459*.

Sun, X.; Li, X.; Li, J.; Wu, F.; Guo, S.; Zhang, T.; and Wang, G. 2023. Text Classification via Large Language Models. *arXiv preprint arXiv:2305.08377*.

Tavangarian, D.; et al. 2004. Is e-Learning the Solution for Individual Learning? *Electronic Journal of E-learning*, 2(2): pp265–272.