Qeios

Peer Review

Review of: "Fine-Tuning PHI-3 for Multiple-Choice Question Answering: Methodology, Results, and Challenges"

Zhongzhou Chen[1]

1. Physics, University of Central Florida, United States

My rating just reflects my evaluation of the current state of the manuscript, which I find to be lacking many significant technical details, and also the argument for the value of the work can be significantly improved. When those details are added, I will most likely change my rating.

First, technical details:

1. I would like to see a much more detailed introduction to the TruthfulQA dataset, hopefully with some example MC questions. More importantly, how exactly are the number of wrong answers retained? This directly affects the possibility of random guessing. Are the answer choices randomized?

2. I have no idea how the prompts are being designed by just reading a single paragraph of text. Please be a lot more detailed on which prompts work best, and show us some examples of your prompts so that others can build their work on yours.

3. I'm no expert in Fine-Tuning, so I'll let an expert in the field comment on whether the fine-tuning details are sufficient. My question is whether fine-tuning was conducted on the entire TruthfulQA dataset or not? If it is fine-tuned on the entire dataset, then maybe the model just learned to answer those specific MCQ questions (just like a student memorizes the answers to a leaked exam)?

4. Please provide the definition of each of the evaluation metrics used for readers who might not be familiar with those metrics. The baseline measure seems to be GPT-3, which is quite an old LLM. I would be more interested in seeing a comparison with either more capable and current LLMs such as GPT-4o, or similar small models. I think it would also be valuable if the author could mention similar performance measures in other related works, especially on the same open dataset.

doi.org/10.32388/K1MJ3R

5. The links for the dataset and full code are all broken, so I can't see them. (This really is a major reason for the one star, because I can't verify if the author actually did the work).

Second, on making the argument for the value of the current study. How does training a very small and compact language model to correctly answer multiple-choice questions improve education? You made the claim in the paper, but I don't think I'm convinced by the arguments that you've provided. What are some potential use cases? What is the advantage of fine-tuning small models over just using the large models? If the end goal is to provide students with feedback and suggestions for study, why not just give the model the correct answer in the prompt? Why should we test if the model can answer the questions correctly themselves? I think the key question here is, who will find this research interesting and valuable, and could build on this study?

I do hope that the lit-review can be more than "so and so did this," but connect this work with related work and point out how the contribution of this work is valuable and unique.

Finally, some comments on the figures: I don't think Figure 1 is needed, since it doesn't show anything relevant to the study, and it doesn't even have axis labels, so I completely don't understand it.

Figure 2 also adds very little value since the process could be easily explained in words. It is also inconsistent in that "Repeat evaluation" is a box, not a loop, while "refine prompt" is in a loop. I also don't see the criteria for satisfactory or unsatisfactory anywhere in this paper.

I hope those comments are constructive, and I'm really interested in seeing the improvements to the paper.

Declarations

Potential competing interests: No potential competing interests to declare.