CSE 584 Final Project

(Yucheng Zhou)

1. Introduction

Nowadays, Large Language Models, such as GPT, Claude, and Llama are showing powerful performances on various tasks like question answering, translation, problem solving, and so on. With the development of recent omni models, more types of input and output can be included, such as audio, image, and video, making them more powerful for multiple tasks. However, it is shown that LLMs may be attacked and give faulty answers. To be more specific, they may try to give answers to problems that are inherently or logically incorrect. One reason for that can be that the model is still not powerful enough to give accurate and confident answers. In this project, we will research on how LLM, specifically, ChatGPT here, performs on the inherently faulty questions.

2. Dataset

In this project, as it is difficult to find related open source dataset on the internet, we come up with the scientific questions and input those questions into ChatGPT to observe and record how they are answered by the LLM. In this project, the model behind our ChatGPT query is the GPT-4o mini, due to the limit of our financial conditions, more advanced models are not explored. In this project, we created a set of 25 query-answer data in total. Besides the faulty science questions and the answer provided by the model, we also include other entries for each data point, such as the discipline of the question, the reason we think it is faulty. In total, our data comes from four science domains, which are Life Science, Geology, Physics, and Mathematics. And for the answers, for simplicity, we didn’t include the whole answer returned by the model, instead, we record the final answer, if it is provided.

3. Research Questions

In this section, we will explore different research questions on the faulty question data.

1) Firstly, we want to explore whether the ChatGPT is indeed fooled by our questions or not, and what is the percentage that it is fooled.

   Asking whether the model is indeed fooled by our questions is because by the design of our questions, they are inherently inconsistent, but there is chance that the model can capture this and directly pointed out to us that the questions are faulty and refuse to give any concrete answer. According to the answers from the dataset, we discovered that 4 out of 25, i.e. 16% of the questions are captured by ChatGPT that they are inherently incorrect. Therefore, 84% of the faulty questions can fool the LLM.

2) Secondly, we want to explore how the LLM performs on each of the subject we provided.

This question is of interest since ChatGPT may be trained differently on different science subject and perform differently. According to the responses, we discovered the following:

In Life Science, 83% of the faulty questions can fool the model.

In Geology, 75% of the faulty questions can fool the model.

In Physics, all the faulty questions can fool the model.

In Math, 50% of the faulty questions can fool the model.

3) Thirdly, we want to explore how can we make the LLM produce more reliable answers.

   To make it possible to produce more reliable answers, we append a prompt after each question “If the problem is not correct itself, you may not answer it.”. In this way, we anticipate that the model can capture some faulty questions. After experiments, we indeed discovered that it recognized two more faulty questions. Now 76% of the faulty questions can fool the LLM. We discovered that after this change, the model is still be fooled by questions that involve more calculations especially in Physics and Math. And it is speculated that the model pays too much attention to the calculation itself due to its relatively low solving ability.

4. Conclusion and future work

To summarize, in this project, we explore 3 different research questions on whether the LLM can be fooled by faulty science questions in section 3. But there is still interesting directions for future research. We can get more data, here 25 data points is a bit less. Second, we may try more models and compare between different models such as GPT4o.