CSE 584 – Fall 2024 – Final Project

Sandhya Somasundaram
939157171
sms9723@psu.edu

December 6 2024, Friday

1 Introduction

Recent advancements in large language models (LLMs) have demonstrated remarkable capabilities in understanding and generating human-like responses across various domains. However, these models often struggle with questions that contain logical inconsistencies or structural flaws. In this project, we explore the limitations of LLMs in identifying and reasoning through faulty science questions. By curating a diverse dataset of intentionally flawed questions across disciplines in Science, we aim to benchmark the reasoning abilities of state-of-the-art LLMs, including ChatGPT, Gemini, and Claude. The primary objective is to assess the models' ability to not only detect faults in these questions but also to provide coherent and logical explanations for their conclusions. We will also go through various research questions and how experimenting with them would help us understand the working of the LLM to enhance it further.

2 Dataset

The dataset for this project was meticulously constructed to ensure diversity in question types, logical flaws, and subject matter. Faulty questions were designed or selected to encompass a variety of disciplines that includes - Chemistry, Physics, Zoology, Geology, Botany, Biology&Genetics, Math, Electricity&Magnetism. Each question was paired with a clear reason explaining why it is flawed. Contributions were sourced three multiple LLMs, including ChatGPT, Gemini, and Claude, to evaluate model-specific performance. It contains 165 entries spanning multiple disciplines. The dataset is organized into five key columns: Discipline, Question, Reason You Think It Is Faulty, Top LLM Used, and Response by the Top LLM. The Discipline column categorizes the questions by domain, enabling targeted analysis across fields. The Question column presents the faulty science questions, crafted to include logical inconsistencies, ambiguous phrasing, or scientifically impossible premises. Each question is accompanied by a detailed explanation in the Reason You Think It Is Faulty column, which outlines the specific flaw, such as contradictions or violations of scientific principles.

3 Research Questions and experiments

3.1 Q1: Performance of LLMs on Multiple-Choice Questions (MCQs) with and without Hints

Experiments & Observations: In the dataset there are about 30MCQ, with 20 being from Biology&Genetics discipline. For the MCQs in the dataset, I included an additional prompt asking the LLM to tell "None" if none of the options were correct. This was tested on ChatGPT, where it successfully identified faulty questions with an accuracy of 58%. For the remaining questions, it still did not identify that the question was faulty. In Gemini, when no prompt was given, only 20% of the questions were identified as faulty. In Claude, all questions were flagged as faulty when no hint was provided. However, when the "none of the options are correct" prompt was given, Claude correctly updated all answers to "None."

Additionally, I tested ChatGPT by providing each question individually and offering valid hints to assess if the right options were being chosen. I recorded responses for approximately 30 MCQs, with around 20 from the Genetics group. On average, five additional prompts were required for each question to provide context and explain why none of the options were correct. After these additional prompts, ChatGPT consistently identified the question as invalid. The results indicate that even when instructed to disregard all options, the LLM still struggled with MCQs. When asked to explain why a question was faulty, the model responded, "I wrongly suggested it as the 'least accurate' without fully considering that all the options are scientifically valid."

These observations highlight that MCQs are particularly challenging for LLMs to comprehend and reason through effectively.

3.2 Q2: Performance of LLMs on MCQs When Given Individually vs. as a Batch

Experiments & Observations: The experiment was conducted in two different ways. In one approach, each question was provided to the LLM individually, and the answer was recorded for each one of it. In the second approach, all the questions were given together, and the model was asked to select the correct answers for each of the questions (i.e. 20 questions were given in one message). The dataset included MCQs from various disciplines, with the majority coming from the Biology&Genetics field, where all questions were MCQs.

For the Biology&Genetics MCQ dataset, which was tested on ChatGPT, the faulty questions were not identified in either of the approaches. The same questions were also tested in Gemini. When presented individually, Gemini identified 60% of the questions as faulty, though the remaining questions went undetected. However, when the questions were given as a batch, only 20% of the questions were identified as invalid. This result suggests that the way questions are presented (individually vs. in a batch) significantly impacts the model's ability to detect faults.

These results were consistent when the prompt to indicate "None" if none of the options were correct was given. In Claude, when the questions were presented as a batch, none were identified as faulty. However, when the questions were given individually, about 80% of the faulty questions were detected. Again, when provided in a batch, Claude failed to recognize any faults.

This indicates that batch processing poses significant challenges for LLMs, especially when it comes to detecting faulty questions. Interestingly, this issue wasn't limited to MCQs alone, even normal descriptive

questions presented in a batch were also frequently not identified as faulty.

3.3 Q3: Performance of Faulty Questions Across Different LLMs and Their Outputs

Experiments & Observations: The dataset used for this experiment contains questions that are faulty in at least one of the tested LLMs. To assess the performance of each LLM, I ran all the questions through three models - ChatGPT, Gemini, and Claude - to evaluate how well each identified faults.

Among the three models, Claude performed the best, identifying about 87% of the questions as faulty. Most of the inaccuracies were found in MCQs. Claude was able to detect faults in other types of questions very well. However, when processing MCQs in a batch (i.e., giving Claude 20 MCQs and asking it to generate the answers), the model struggled to perform well, indicating that batch processing posed challenges for fault detection in MCQs.

I then tried the same with Gemini and ChatGPT. In comparison, Gemini performed better than ChatGPT. In ChatGPT, approximately 77% of the questions went undetected, meaning that the model failed to identify the faults in those questions. On the other hand, Gemini identified faults in 52% of the questions, leaving only 48% undetected.

These results highlight that the architecture and design of each LLM have a significant impact on its accuracy in fault detection. While Claude performed well in identifying faulty questions, the batch processing issue affected its performance. Gemini performed reasonably well, but ChatGPT struggled significantly with detecting faults in the dataset.

3.4 Q4: Does the Framing of a Question Affect the Response of the LLM?

Experiments & Observations: Consider the question: "Explain how deforestation simultaneously increases rainfall and reduces water availability in a single ecosystem." This question contains a fundamental flaw, as deforestation does not increase rainfall in the long run and is generally harmful to the environment. However, when presented to the LLMs(GPTs), the models generate responses instead of recognizing the flaw.

Interestingly, when the question is reframed as "Does deforestation increase rainfall?", the response from GPT is correctly a "NO", indicating that the way the question is framed can significantly affect the model's response.

Additionally, when asked, "How does deforestation increase rainfall?", the flaw goes undetected in GPT, which proceeds to provide an answer. However, both Gemini and Claude successfully identify the flaw in this question.

These findings demonstrate that the framing of a question plays a crucial role in how LLMs process and respond to it, with different models varying in their ability to detect underlying flaws based on question structure.

3.5 Q5: Consistency of LLM Responses: How Consistent Are LLMs in Identifying Faulty Results?

Experiments & Observations: I ran the same set of faulty data four times through each of the three LLMs - ChatGPT, Gemini, and Claude - to evaluate how consistent their responses were in identifying flaws.

In Claude, only 20% of the time the faults were not identifoied in the question, and this occurred exclusively with MCQs. However, after the first run, the flaws were consistently identified in subsequent repetitions.

For Gemini, flaws were identified 60% of the time by the fourth repetition. No extra prompts were given, and the same question was asked four times. This demonstrated a moderate level of consistency, where the model gradually improved in identifying faults over multiple runs.

In contrast, ChatGPT showed the most consistency. About 80% of the time, faulty questions were not identified, and even after repeating the same question, the model continued to provide the same incorrect response.

These findings suggest that while some models, like Claude, quickly adapt and improve in detecting flaws, others, like ChatGPT, significantly lack in identifying faults, even after multiple attempts.