CSE584 Final Project Report

Faulty Science Questions

Mega Sri Shyam B
mpb6512@psu.edu

1. Introduction

In recent years, large language models (LLMs) such as ChatGPT, GPT-4, Claude-3-Opus, Gemini-1.5-Pro, Bard, and LLaMA have demonstrated remarkable capabilities in understanding and generating human-like text across various domains. These models are employed in diverse applications, including education, research, customer service, and more. However, despite their advanced proficiency, LLMs are not infallible and can be susceptible to producing incorrect or misleading answers, particularly when confronted with poorly structured or inherently flawed questions.

This project aims to investigate the robustness of top-performing LLMs against **faulty science questions**—questions intentionally designed with logical flaws, ambiguities, or misleading information to trick the models into providing incorrect answers. By compiling a comprehensive dataset of such questions across multiple disciplines (Mathematics, Physics, Chemistry, Biology, Astronomy), this study seeks to evaluate the ability of various LLMs to recognize and appropriately handle these faulty inputs. Understanding the limitations of LLMs in this context is crucial for improving their reliability and ensuring their effective deployment in sensitive applications.

2. Dataset Overview

The dataset comprises **100 faulty science questions** distributed across **10 disciplines**, with each discipline containing **10 unique questions**. These questions are meticulously crafted to include subtle flaws or ambiguities that can mislead LLMs. The dataset is structured in a Google Sheet format with five columns:

1. **Discipline**: The subject area of the question (e.g., Mathematics, Physics).
2. **Question**: The faulty science question designed to trick the LLM.
3. **Reason the Question is Faulty**: An explanation of the inherent flaw or ambiguity in the question.
4. **Which Top LLM You Tried**: The name of the LLM used to test the question (e.g., ChatGPT, GPT-4).
5. **Response by a Top LLM**: The answer provided by the LLM to the faulty question.

3. Research Questions

Based on the dataset, the following research questions have been formulated to guide the experimental analysis:

1. **RQ1: Can LLMs correctly identify and reject faulty science questions?**
   - *Objective*: To determine whether models like GPT-4, ChatGPT, and Claude-3 can recognize that a question is inherently flawed or misleading.
2. **RQ2: What types of faults in the questions (ambiguity, unrealistic assumptions, etc.) lead to the most frequent incorrect answers?**
   - *Objective*: To analyze which specific flaws (e.g., word problem ambiguities, logical fallacies) are most likely to cause LLMs to provide incorrect responses.
3. **RQ3: Do different LLMs (ChatGPT, GPT-4, Claude-3, etc.) exhibit similar or different behaviors when answering faulty questions?**
   - *Objective*: To compare the performance and error patterns of various LLMs when confronted with the same faulty questions.
4. **RQ4: How do LLMs handle word problems involving unrealistic scenarios, such as negative quantities or impossible real-world conditions?**
   - *Objective*: To evaluate whether LLMs can discern and appropriately respond to questions that present logically impossible scenarios.
5. **RQ5: What is the impact of question phrasing on LLM performance? Does vague or misleading phrasing lead to consistent failures?**
   - *Objective*: To investigate how subtle changes in the wording and structure of a question influence the ability of LLMs to generate correct answers.

4. Experiment Design

To address the research questions, a systematic experimental approach was adopted, involving the evaluation of responses from seven top-performing LLMs. The following outlines the experimental setup and methodology:

Participants:

- The participants in this study are the seven LLMs:
  - ChatGPT
  - GPT-4
  - Claude-3-Opus
  - Gemini-1.5-Pro
  - Bard
  - LLaMA

Procedure:

1. **Selection of Questions**: From the dataset, 20 faulty math questions were selected, ensuring a variety of flaws (e.g., negative results, ambiguous phrasing).
2. **Question Submission**: Each faulty question was submitted to all seven LLMs individually.

3. **Response Recording**: The responses from each LLM were recorded without any prior indication that the questions were faulty.
4. **Classification of Responses**: Responses were categorized as:
    - **Correct**: The LLM correctly identifies the flaw and provides an appropriate response or correction.
    - **Incorrect**: The LLM provides an answer based on the flawed premise without recognizing the error.
    - **Failed to Answer**: The LLM does not provide a coherent answer or fails to address the question.

Metrics:

- **Accuracy Rate**: The percentage of correct responses out of total attempts.
- **Error Analysis**: Identification of common error types (logical fallacies, computational mistakes, misinterpretation of units).
- **Cross-LLM Comparison**: Evaluation of performance differences among the LLMs.

Example Experiment:

- **Faulty Question**: "Lily received 3 cookies from her best friend yesterday and ate 5 for breakfast. Today, her friend gave her 3 more cookies. How many cookies does Lily have now?"
    - **Expected Fault**: The question implies a negative number of cookies, which is illogical.
    - **Objective**: Assess whether the LLM recognizes the flaw or simply computes the arithmetic result.

5. Results and Analysis

RQ1: Can LLMs correctly identify and reject faulty science questions?

- **Findings**:
    - **ChatGPT**: Frequently attempted to compute the arithmetic result without recognizing the flaw.
    - **GPT-4**: Similar behavior to ChatGPT, though with slightly better contextual understanding.
    - **Claude-3-Opus**: Mixed responses; some recognition of logical inconsistency but often provided numerical answers.
    - **Gemini-1.5-Pro**: Showed limited ability to detect faulty premises, focusing on computation.
    - **Bard**: Consistently provided arithmetic results, rarely addressing the logical flaw.
    - **LLaMA**: Predominantly performed calculations without recognizing the issue.
- **Conclusion**: Most LLMs do not inherently recognize faulty questions and proceed to compute based on the given data, resulting in incorrect or illogical answers.

RQ2: What types of faults in the questions lead to the most frequent incorrect answers?

- **Findings**:

  - **Negative Quantities**: Questions implying negative real-world quantities (e.g., negative cookies) led to incorrect answers as models computed numerically.
  - **Ambiguous Phrasing**: Questions with unclear instructions or missing context resulted in misinterpretations and wrong computations.
  - **Unit Confusion**: Mixed units without clear conversion instructions caused models to apply incorrect conversion factors or ignore the issue.
  - **Logical Fallacies**: Scenarios presenting impossible real-world conditions (e.g., giving away more items than possessed) were not identified as flawed by most models.
- **Conclusion**: Ambiguities and unrealistic assumptions are primary drivers of incorrect responses in faulty questions.

RQ3: Do different LLMs exhibit similar or different behaviors when answering faulty questions?

- **Findings**:
  - **Similarities**: All models, except Anthropic, largely ignored the faulty nature and computed based on provided numbers.
  - **Differences**: Anthropic occasionally recognized logical inconsistencies and provided more context-aware responses. GPT-4 showed marginal improvements in contextual understanding over ChatGPT.
- **Conclusion**: While there are overarching similarities in behavior, certain models like Anthropic exhibit better contextual handling, suggesting potential architectural or training differences.

RQ4: How do LLMs handle word problems involving unrealistic scenarios, such as negative quantities or impossible real-world conditions?

- **Findings**:
  - **Consistent Computation**: Most LLMs computed negative results when prompted with unrealistic scenarios without flagging the illogical aspect.
  - **Lack of Real-World Reasoning**: Models did not apply real-world constraints that would make the scenario impossible, focusing solely on mathematical computation.
- **Conclusion**: LLMs lack the nuanced real-world reasoning necessary to identify and appropriately respond to unrealistic scenarios within word problems.

RQ5: What is the impact of question phrasing on LLM performance? Does vague or misleading phrasing lead to consistent failures?

- **Findings**:
  - **Vague Phrasing**: Questions with vague or unclear instructions led to consistent computational errors as models misinterpreted the intended meaning.
  - **Misleading Instructions**: Questions structured to imply certain logical flaws without explicit indicators resulted in models overlooking the inherent issues and providing incorrect answers.
- **Conclusion**: Subtle variations in question phrasing significantly impact LLM performance, with vague or misleading wording leading to higher rates of incorrect responses.

6. Conclusion

This project underscores the limitations of current LLMs in handling faulty or misleading science questions. Despite their advanced capabilities, models like ChatGPT, GPT-4, Claude-3-Opus, Gemini-1.5-Pro, Bard, and LLaMA predominantly focus on computational accuracy without assessing the logical coherence of the questions. As a result, they frequently provide incorrect answers when confronted with scenarios that contain inherent flaws or ambiguities.

Key Takeaways:

- **LLMs' Computational Focus**: LLMs excel at performing calculations but lack the ability to discern logical inconsistencies or unrealistic assumptions within questions.
- **Ambiguity and Misleading Phrasing**: Subtle flaws in question phrasing significantly degrade the accuracy of LLM responses, highlighting the need for improved contextual understanding.
- **Model Variability**: While most LLMs exhibit similar shortcomings, some models like Anthropic show potential for better handling of flawed inputs, indicating room for architectural enhancements.