CSE 584 Final Project

Samira Malek

8 December 2024

Abstract

I have developed a set of intentionally flawed science questions in subjects such as mathematics, physics, geography, and linguistics. These questions are designed to be straightforward for humans to answer but reveal errors in the performance of Large Language Models (LLMs). This study analyzes the responses of three prominent LLMs—Llama, Gemini, and GPT-4—to these questions, focusing on their final answers and the reasoning processes behind them. Additionally, we explore strategies to improve or correct LLM responses through refined prompt engineering. The study also investigates whether altering the linguistic structure of the flawed questions affects the models' ability to detect the embedded errors.

1 Introduction

Large Language Models (LLMs) have rapidly gained prominence as powerful tools across various domains, demonstrating remarkable capabilities in tasks such as language understanding, reasoning, and knowledge synthesis. However, their performance often falters in nuanced scenarios, particularly when exposed to subtle flaws in logic or reasoning, raising critical concerns about their reliability in high-stakes or unsupervised applications.

Building upon these observations, our project investigates the specific vulnerabilities of LLMs by creating and testing a set of faulty science questions spanning mathematics, physics, geography, and linguistics. These questions are deliberately crafted to contain subtle errors that are intuitive for humans to identify but challenging for LLMs to detect. By assessing the responses of three prominent LLMs—Llama, Gemini, and GPT-4—we explore differences in their reasoning processes, final outputs, and susceptibility to such crafted errors.

Beyond merely evaluating model performance, we delve into the interplay between question formulation and model behavior. Through rigorous prompt engineering, we analyze whether rephrasing or altering the linguistic structure of these questions influences the models' ability to identify and address the embedded flaws. Our findings highlight the nuanced relationship between question presentation, model interpretability, and response quality.

This study not only illuminates current limitations in LLM reasoning but also underscores the importance of developing more robust prompt engineering techniques. By leveraging these insights, we aim to enhance the precision and reliability of LLMs, fostering a deeper understanding of their capabilities and paving the way for more informed and effective human-LLM collaboration.

2 Prior Works

Large Language Models (LLMs) have demonstrated significant challenges in effectively interpreting nuanced human language, often misinterpreting idiomatic expressions and failing to grasp subtle linguistic intricacies. This limitation, rooted in their intrinsic operational mechanisms, hampers their ability to handle tasks requiring deep linguistic comprehension . Similarly, LLMs struggle with common sense reasoning due to their disembodied nature, as they lack the sensory experiences essential for developing human-like intuition. This absence of embodied interaction, a concept emphasized by Hubert Dreyfus [2], restricts their capacity to reason through scenarios that rely on experiential understanding .

Contextual reasoning is another area where LLMs face difficulties. They often fail to process implicit contextual cues, a critical aspect of accurate interpretation and decision-making . This shortcoming underscores a persistent challenge in achieving human-like comprehension of complex, situational nuances. Furthermore, LLMs lack proficiency in visual-spatial reasoning, as they are unable to mentally represent or manipulate objects in space effectively [4]. Their inability to understand spatial relationships limits their application in tasks requiring geometric or physical reasoning.

Despite advancements, LLMs exhibit fragility in mathematical reasoning, especially in basic arithmetic or word-based problems. These models lack an inherent numerical framework and often depend on external computation tools to achieve accuracy [1]. Additionally, they are prone to reproducing inaccuracies embedded within their training data, particularly in popular science domains, where outdated or erroneous concepts are perpetuated [5]. This limitation highlights the critical need for curating and updating training datasets to enhance the reliability of their outputs.

Relational reasoning, involving temporal, causal, or other connections between entities, also remains a significant hurdle for LLMs. Their interpretations frequently lack the depth necessary for solving problems requiring an intuitive grasp of relationships. Similarly, logical reasoning capabilities in LLMs, while promising in simulating certain aspects, fall short of delivering consistent and reliable human-like reasoning. Research has shown mixed results in their ability to solve logical problems, underscoring the need for continued refinement [3].

Overfitting further exacerbates these issues, as LLMs often excel within the confines of their training data but struggle to generalize effectively to novel scenarios. This lack of robustness in extrapolation tasks underscores the broader challenges of achieving true adaptability in these models. Finally, while benchmarks provide a standardized method of evaluating LLM performance, their ability to reflect real-world capabilities has been called into question. Many benchmarks prioritize optimization for rankings over holistic performance, potentially misrepresenting the true strengths and limitations of these models [6].

These limitations form the foundation of our work, which seeks to systematically investigate LLM deficiencies using a curated benchmark of faulty science questions across various domains. By addressing weaknesses in linguistic, mathematical, and logical reasoning, as well as their susceptibility to contextual and commonsense errors, this study aims to explore not only the current capabilities of LLMs but also avenues for improvement through refined prompt engineering and innovative question design.

3 Models

In this project, we utilized three state-of-the-art Large Language Models (LLMs) to evaluate their performance and reasoning capabilities across various domains: **Llama**, **Gemini**, and **ChatGPT**. These models were chosen to represent a diverse range of architectures and capabilities, allowing for a comprehensive analysis of their strengths and limitations in handling linguistically and conceptually challenging tasks. Table 1 summarizes the specific versions of the models used in this study.

1. **Llama (Llama 3.1 70b Turbo):**
   Llama is an advanced language model known for its scalable architecture and efficiency in natural language processing tasks. The version used, Llama 3.1 70b Turbo, is equipped with 70 billion parameters and optimized for faster inference, making it a suitable candidate for testing both linguistic and mathematical reasoning capabilities. Its large parameter count allows it to capture intricate patterns in text, although its performance on nuanced tasks is still under exploration.

2. **Gemini (Gemini 1.5 with Flash):**
   Gemini is a cutting-edge LLM that integrates enhanced contextual understanding and reasoning capabilities. The version used in this project, Gemini 1.5 with Flash, incorporates optimizations designed to improve processing speed and contextual comprehension. This model has been recognized for its advancements in maintaining coherence across longer text sequences, making it an interesting comparison point for evaluating responses to complex, context-dependent questions.

3. **ChatGPT (GPT 4o):**
   ChatGPT, based on the GPT-4 architecture, represents one of the most widely adopted LLMs for general-purpose applications. The version employed in this study, GPT 4o, is a lightweight variant aimed at balancing performance and efficiency. Known for its robust reasoning and general knowledge capabilities, GPT 4o serves as a benchmark for comparing the reasoning accuracy and response quality of the other models.

Each model was tested using a curated set of faulty science questions designed to probe their reasoning, linguistic understanding, and ability to detect logical inconsistencies. By analyzing the performance of these models, this study provides insights into their respective capabilities, limitations, and potential avenues for refinement through improved prompt engineering and question design. The diversity in model architectures and optimization strategies offers a rich context for evaluating how different LLMs approach the same set of challenges.

Model	**Model Version**
Llama	Llama 3.1 70b Turbo
Gemini	Gemini 1.5 with Flash
ChatGPT	GPT 4o

Table 1: Models and Model Versions which are used in this project.

4 Criteria to Compute Accuracy

To evaluate the performance of the Large Language Models (LLMs) in this study, we established a structured framework for assessing their responses. The evaluation is based on two key aspects: the correctness of the **Final Answer** and the accuracy of the **Thought Process** leading to that answer. Each aspect is scored based on predefined criteria, as outlined in Table 2.

4.1 Evaluation Criteria

- **Final Answer**: The correctness of the final response provided by the LLM is a critical factor in determining its overall accuracy. This criterion is binary, with the following scoring levels:
  - **Correct**: If the final answer aligns with the expected solution, the model is awarded 100%.
  - **False**: If the final answer is incorrect or fails to address the question appropriately, the model receives 0%.
- **Thought Process**: This criterion evaluates the reasoning and logical steps taken by the LLM to arrive at its final answer. It accounts for the coherence, relevance, and correctness of the intermediate steps, even if the final answer is incorrect. The scoring levels are:
  - **Correct**: If the reasoning is entirely logical, consistent, and accurate, 100% is awarded.
  - **Somewhat Correct**: If the thought process contains partial correctness or logical gaps but demonstrates an attempt to reason through the problem, a score of 50% is given.
  - **Wrong**: If the reasoning is fundamentally flawed, irrelevant, or illogical, the model is awarded 0%.

4.2 Scoring Approach

Each response is independently scored for the **Final Answer** and **Thought Process**. This dual-layered evaluation provides a nuanced understanding of the models' reasoning capabilities, distinguishing between cases where they fail due to incorrect final answers but exhibit sound reasoning, and those where both the reasoning and answers are incorrect. By integrating these criteria, the study offers a comprehensive assessment of LLM performance, highlighting areas for improvement and refinement.

This approach ensures that both the outcome and the underlying reasoning are rigorously analyzed, providing deeper insights into the models' capabilities and limitations.

Criteria	Criteria Level	Points
Final Answer	Correct	100%
Final Answer	False	0%
Thought Process	Correct	100%
Thought Process	Somewhat Correct	50%
Thought Process	Wrong	0%

Table 2: Criteria for computing the accuracy of responses.

5 Chang Prompt to correct LLMs Output

To enhance the performance of Large Language Models (LLMs) and address their susceptibility to faulty questions, we introduced a specific modification to the prompt. The goal of this approach is to guide the models toward verifying the logical correctness of a question before attempting to answer it. This additional step encourages the models to identify any inconsistencies or errors in the question itself, thereby improving both the accuracy of the final answers and the reasoning process. The following prompt modification was applied:

Prompt text template

Faulty Science Question + **"check the correctness of the question, then if the question is correct, find the answer."**.

The following is an example of a faulty science question designed to evaluate the effectiveness of this prompt modification:

Prompt text example

Bob participated in an exam with 100 questions. Each correct response is worth 5 points and each false response has -2 points. He correctly answered 85 questions and he gave wrong answer to 16 questions. What is Bob's total score?
check the correctness of the question, then if the question is correct, find the answer.

5.1 Model-Specific Observations

GPT-4o: As shown in Figure 1, GPT-4o successfully identified the fault in the question after applying the modified prompt and provided a logical explanation of the inconsistency. This demonstrates that the added instruction significantly improved its reasoning process and final answer accuracy.
Llama: Figure 2 illustrates how the modified prompt enabled Llama to detect the faulty logic in the question. This result highlights Llama's improved ability to reason through the question's structure before attempting to answer it.
Gemini: In contrast, Figure 3 shows that the modified prompt did not significantly help Gemini. While there was a slight improvement in its thought process, it still failed to identify

the inconsistency or provide an accurate final answer.

Let’s verify the correctness of the question first:

1. Bob participated in an exam with 100 questions.
    - Bob correctly answered 85 questions.
    - Bob gave wrong answers to 16 questions.
    - Total questions answered by Bob: $85 + 16 = 101$.
    - This exceeds 100, so the question contains an error.

Since the number of questions answered exceeds the total number of questions in the exam, the question is **incorrect**.

If you want me to modify the question or proceed with specific assumptions, let me know!

Figure 1: The output of GPT-4o for prompt text example.

Llama 3.1 70b Turbo

Let's check the correctness of the question.

The question states that Bob participated in an exam with 100 questions. It also mentions that Bob correctly answered 85 questions and gave a wrong answer to 16 questions.

However, the total number of questions Bob answered (either correctly or incorrectly) is 85 + 16 = 101, which exceeds the total number of questions in the exam (100). This means that Bob cannot have answered 85 questions correctly and 16 questions incorrectly in a 100-question exam.

So, the question is incorrect. The numbers don't add up.

If you'd like to revise the question or provide more accurate information, I'd be happy to help you find the answer.

Figure 2: The output of Llama for prompt text example.

The question is correct.

To find Bob's total score, we first calculate the points he earned from the correct answers and the points he lost from the incorrect answers.

Points earned from correct answers = 85 questions * 5 points/question = 425 points
Points lost from incorrect answers = 16 questions * 2 points/question = 32 points

Now, we subtract the points lost from the points earned:

Total score = Points earned - Points lost = 425 points - 32 points = 393 points

Therefore, Bob's total score is 393 points.

Figure 3: The output of Gemini for prompt text example.

5.2 Effectiveness of the Proposed Method

The results summarized in Table 3 demonstrate that the modified prompt effectively increased both the final answer accuracy and the thought process accuracy for most models. Llama and GPT-4o showed substantial improvements in their ability to reason through faulty questions, with Llama achieving an 80% accuracy in both final answers and thought processes. GPT-4o achieved a 60% accuracy in both metrics. Gemini, however, showed minimal improvement, indicating potential limitations in its current architecture or training methodology.

Model	Final Answer Accuracy	Thought Process Accuracy
Llama	80%	80%
Gemini	0%	10%
ChatGPT	60%	60%

Table 3: Final answer and thought process accuracy for all three models.

6 Conclusion

The results confirm the effectiveness of the proposed prompt modification. By explicitly instructing models to verify the correctness of the question before solving it, we observed significant improvements in reasoning and final answers, particularly for Llama and GPT-4o. This approach highlights the importance of prompt engineering as a tool for addressing LLM limitations and guiding them towards more accurate and logical outputs. The findings underscore the potential of this method for enhancing the performance of LLMs in complex reasoning tasks.

References

[1] Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. Large language models for mathematical reasoning: Progresses and challenges. *arXiv preprint arXiv:2402.00157*, 2024.

[2] Nicholas Asher, Swarnadeep Bhar, Akshay Chaturvedi, Julie Hunter, and Soumya Paul. Limits for learning with language models. *arXiv preprint arXiv:2306.12213*, 2023.

[3] Panagiotis Giadikiaroglou, Maria Lymperaiou, Giorgos Filandrianos, and Giorgos Stamou. Puzzle solving using reasoning of large language models: A survey. *arXiv preprint arXiv:2402.11291*, 2024.

[4] Miyu Sasaki, Natsumi Watanabe, and Tsukihito Komanaka. Enhancing contextual understanding of mistral llm with external knowledge bases. 2024.

[5] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Do large language models have a legal duty to tell the truth? *Royal Society Open Science*, 11(8):240197, 2024.

[6] Sean Williams and James Huckle. Easy problems that llms get wrong. *arXiv preprint arXiv:2405.19616*, 2024.