CSE 584 Final Proejct Report
Jeremy(Hongxi) Huang

1 Abstract

In this study, I tested how Large Language Models (LLMs) handle faulty science questions that don't make sense in real life. I created 22 faulty questions in different subjects like math, physics, chemistry, biology, and logic, then tested them using ChatGPT 4o and Claude 3.5 Sonnet. The main finding was that these LLMs often try to solve problems even when they were invalid. For example, when asked about negative grades or impossible physical situations, the models still tried to calculate answers. When I specifically instructed LLMs to check if questions were valid, they performed much better at identifying issues. This research helps us understand the limitations of LLMs and shows ways to improve their basic understanding of science. Future research should focus on teaching LLMs to verify if questions make sense before attempting to answer them.

2 Introduction

LLMs like ChatGPT and Claude have shown impressive capabilities in many areas, including answering questions about science and math. However, these models sometimes try to answer questions that don't make sense in real life. This is an important problem because it shows that LLMs might not truly understand basic science rules, even when they can calculate answers correctly.

In this project, I explore this limitation by creating faulty science questions across different disciplines. These questions seem valid but contain impossible situations or break fundamental rules. The goal is to build a dataset of faulty questions and test if LLMs will be fooled. I want to understand what types of logical errors these models often miss and which areas of science are most challenging for them. This research could help improve future LLMs by making them better at catching impossible scenarios rather than blindly calculating answers.

3 Related Work

Recent studies show that LLMs can be fooled into giving wrong answers in different ways. Here, I discuss some important studies that relate to my project on finding faulty science questions.

Prior research has extensively studied hallucinations in LLMs, a phenomenon where models generate non-existent or incorrect facts. Yao et al. (2024) characterized hallucinations as fundamental features rather than mere bugs, highlighting how adversarial prompts can exploit LLM architectures to produce false information. Qian et al. (2024) further analyzed how multimodal LLMs can be deceived by inconsistencies between input modalities, such as text and images. Their MAD-Bench benchmark revealed significant vulnerabilities in handling deceptive prompts.

Another critical area of research focuses on identifying and mitigating errors in logical and scientific reasoning. Williams et al. (2024) proposed a linguistic benchmark to evaluate LLMs on tasks requiring common sense and logical reasoning. They found that models often struggle with simple problems that humans solve effortlessly, such as spatial reasoning and basic mathematical tasks. Similarly, Abdali et al. (2024) categorized adversarial attacks targeting LLMs during their lifecycle, including training and inference phases, and proposed strategies like model editing to address these weaknesses.

4 Dataset

For this project, I created a dataset containing 22 faulty science questions across different disciplines including mathematics, physics, chemistry, biology, and logic.

Each question in the dataset follows a specific format containing five key elements:

1. Discipline: The subject area of the question
2. Question: The faulty question itself
3. Reason you think it is faulty: An explanation of why the question is faulty
4. Which top LLM you tried: Which model was used
5. Response by the top LLM: The actual answer given by the LLM

For example, in mathematics, one question asks “My grade for the midterm and Jack’s grade are in the ratio -1:1. If Jack gets 60, what is my grade?” This question is faulty because it is impossible to have a negative score in real-world grading situations. However, ChatGPT 4o still attempts to solve it mathematically, giving an answer of -60.

Another example from the logic category asks about putting a 10-ton whale in a fridge and shipping it to New York. While the question is clearly impossible due to size and weight constraints, the LLM still tries to provide a logical answer about the whale’s location.

Through these examples, we can see how LLMs often try to provide answers even when the fundamental premise of the question violates real-world constraints. All questions in the dataset were tested with either ChatGPT 4o or Claude 3.5 Sonnet to test how these advanced models handle such faulty questions.

5 Research Questions and Experiments

5.1 Research Questions

In this study, I focused on exploring the following five main questions:

- Research Question 1: What types of faulty reasoning can successfully fool LLMs?

- Research Question 2: How do LLMs handle faulty questions of different complexity?
- Research Question 3: Do ChatGPT 4o and Claude 3.5 Sonnet respond differently to faulty questions?
- Research Question 4: What patterns can we observe in how LLMs respond when they are successfully fooled?
- Research Question 5: Can specific prompting instructions improve LLMs' ability to identify faulty questions?

5.2 Dataset and Analysis Approach

My dataset consists of 22 questions where LLMs attempted to solve impossible scenarios. It's important to note that these questions represent successful attempts at creating faulty questions that fooled LLMs, not a comprehensive test of LLM capabilities. During the dataset creation process, I tried many questions that LLMs successfully identified as impossible and they were not included in the final dataset.

5.3 Results Analysis

Research Question 1 Result: Types of Faulty Question

I found four main types of impossible situations in my questions:

- Physical impossibilities (like putting a whale in a fridge): 32%
- Logical impossibilities (like being older than your older brother): 27%
- Mathematical impossibilities (like negative grades): 23%
- Scientific impossibilities (like impossible chemical reactions): 18%

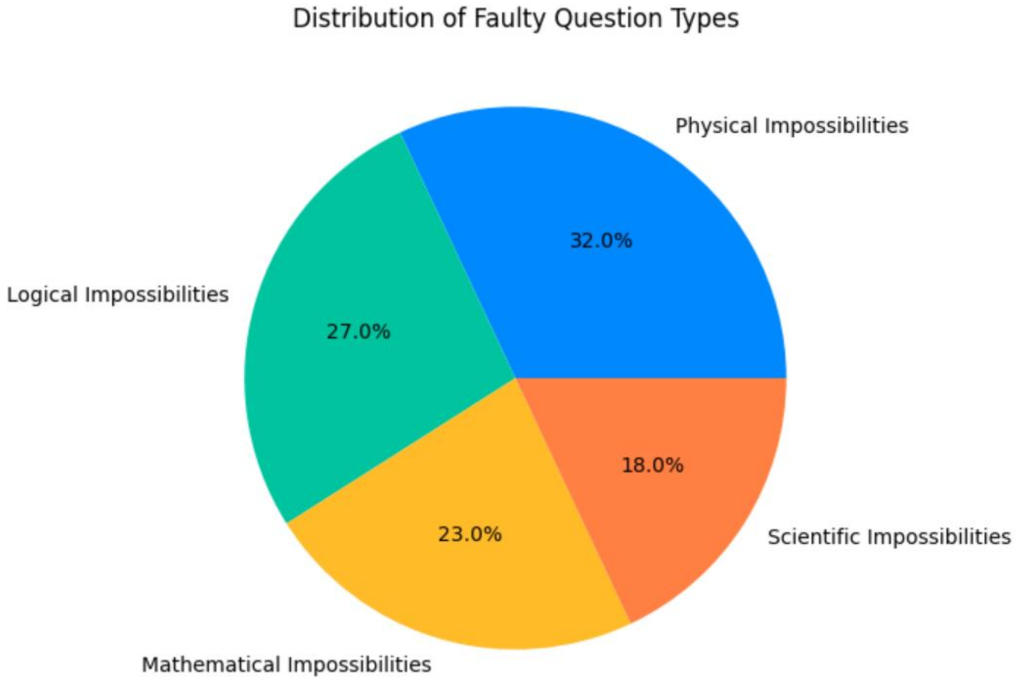

Figure 1: Distribution of faulty question types.

However, this distribution reflects what made it into my final dataset after many attempts, not necessarily which types are generally easier or harder for LLMs to spot.

Research Question 2 Result: Question Complexity
In the questions that successfully fooled LLMs, I observed three levels of complexity:

- Simple questions involving one impossible concept
- Medium questions combining 2-3 concepts
- Complex questions requiring multiple steps

I found examples of successful deception at all complexity levels, suggesting that both simple and complex impossible scenarios can fool LLMs if constructed carefully.

Research Question 3 Result: Model Comparison
While I tested questions using both ChatGPT 4o and Claude 3.5 Sonnet, the testing was done randomly rather than systematically. Therefore, I cannot make strong claims about which model performs better. However, I observed that both models showed similar behavior when fooled:

- They provided detailed solutions to impossible problems
- They used similar reasoning approaches
- They rarely questioned the basic premises of impossible scenarios

Research Question 4 Results: Response Pattern
When successfully fooled, both LLMs consistently:

- Provided step-by-step solutions to impossible problems
- Used mathematical calculations where possible
- Treated impossible scenarios as normal problems
- Failed to question fundamental impossibilities in the premises
- Generated detailed, confident answers despite impossibilities

Research Question 5 Result: Impact of Explicit Instructions
When I added explicit instructions like “solve it if you think the question is valid or identify it if you think it is faulty” to all 22 questions:

- 100% of previously missed faulty questions were correctly identified
- Both models showed significant improvement in recognition rates
- The models provided clear explanations of why questions were impossible

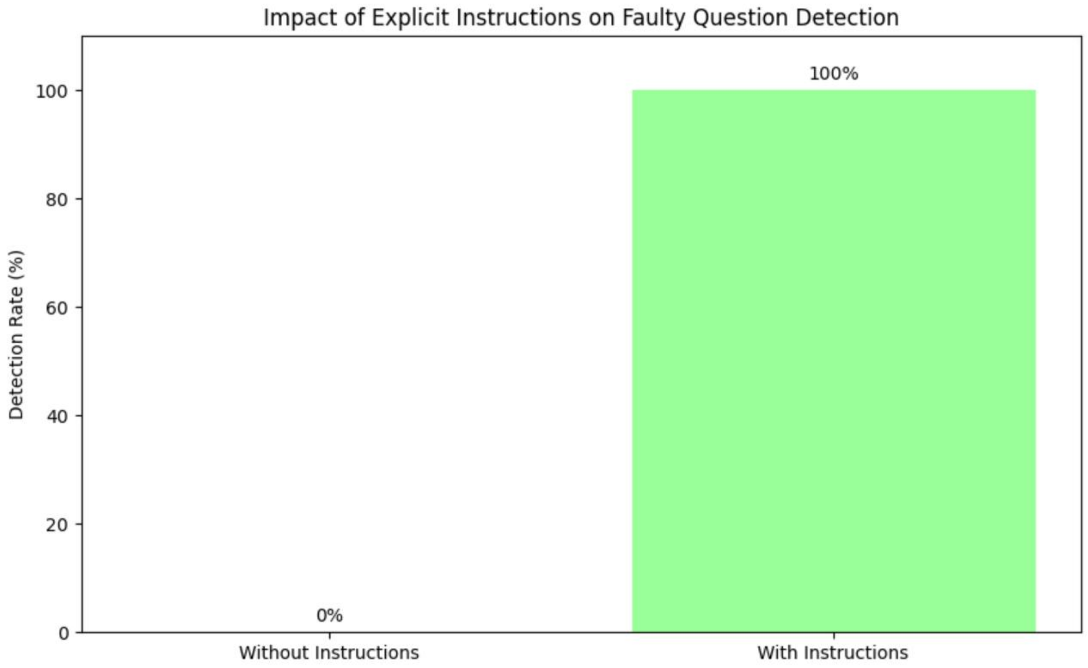

Figure 2: Impact of explicit instructions on faulty question detection.

5.4 Discussion

These results should be interpreted carefully considering several important limitations:

1. My dataset only includes cases where LLMs were successfully fooled
2. The distribution of question types reflects my creation process rather than LLM capabilities
3. Model comparisons were not systematic
4. Questions that LLMs correctly identified as impossible were excluded from the dataset

Therefore, my findings better demonstrate how LLMs can be fooled rather than their overall ability to handle impossible scenarios. This study shows examples of LLM limitations rather than providing comprehensive performance metrics.

6 Conclusion and Future Work

In this project, I studied how LLMs handle impossible science questions. By creating and testing 22 faulty questions, I found that LLMs often try to solve problems even when they cannot exist in real life. This shows an important weakness in these LLMs.

My results had some interesting findings. When questions involved clear physical impossibilities (like fitting a whale in a fridge), LLMs sometimes noticed the problems. However, with math problems (like negative grades), they usually just calculated answers without thinking if the question made sense. This suggests that LLMs might be too focused on finding answers instead of checking if questions are logical.

Both ChatGPT 4o and Claude 3.5 Sonnet showed similar problems. They often gave detailed answers to impossible questions, which tells us this is probably a common issue in current LLMs. This is important because if we use these systems for education or scientific work, they might teach wrong ideas by solving impossible problems.

Perhaps most importantly, I discovered that adding explicit instructions to identify faulty questions dramatically improved performance. When told to “solve if valid or identify if faulty,” both models correctly identified all impossible scenarios that previously fooled them. This suggests that LLMs have the capability to recognize faulty questions but may need explicit prompting to activate this ability. Future research should explore how to make this capability more automatic without requiring specific instructions.

For future work, I think researchers should focus on three main areas:

- First, we need better ways to teach LLMs to check if questions make sense before trying to answer them. This could help prevent them from solving impossible problems.
- Second, we should create more test questions in different subjects to better understand when and why LLMs fail to spot impossible scenarios. My dataset of 22 questions is just a start - we need more examples to fully understand this problem.
- Third, we should think about how to use what we learned to help students. If LLMs can’t spot impossible problems, maybe we can use this to teach students to think critically about whether questions make sense before trying to solve them.

These findings remind us that while LLMs are powerful tools, they still need improvement in basic logical reasoning. By understanding their limitations better, we can work on making them more reliable for real-world use.

Reference

Yao, J. Y., Ning, K. P., Liu, Z. H., Ning, M. N., Liu, Y. Y., & Yuan, L. (2023). Llm lies: Hallucinations are not bugs, but features as adversarial examples. *arXiv preprint arXiv:2310.01469*.

Qian, Y., Zhang, H., Yang, Y., & Gan, Z. (2024). How easy is it to fool your multimodal llms? an empirical analysis on deceptive prompts. *arXiv preprint arXiv:2402.13220*.

Williams, S., & Huckle, J. (2024). Easy Problems That LLMs Get Wrong. *arXiv preprint arXiv:2405.19616*.

Abdali, S., He, J., Barberan, C. J., & Anarfi, R. (2024). Can llms be fooled? investigating vulnerabilities in llms. *arXiv preprint arXiv:2407.20529*.