# Faulty Science Questions for LLM Evaluation

Jiamu Bai

December 5, 2024

## 1 Dataset Description

The dataset contains questions intentionally designed to confuse LLMs, highlighting specific vulnerabilities.

- Disciplines: Currently, the dataset primarily focuses on mathematical problems. Structure: Each entry in the dataset includes:
- Structure: Each entry in the dataset includes:
  - The scientific discipline (Math and Physics).
  - The faulty question.
  - A reason explaining why the question is faulty.
  - The LLM tested (ChatGPT-4).
  - The LLM's response.

The data can be found in the link:
https://github.com/jiamubai/cse-584/blob/main/final_project.md

## 2 Experiment Design

The faulty questions in this dataset were systematically crafted by modifying correct science questions to introduce logical or factual inconsistencies, making them implausible or nonsensical. This process was performed carefully to ensure the faults were subtle enough to potentially mislead LLMs.

I start with correct questions by creating or sourcing accurate and valid questions in disciplines like Mathematics and Physics. Then, I introducing faults deliberately by altering key information in the questions to break logical or physical rules while maintaining surface plausibility.

**Examples of Fault Creation.**

- Faulty: "A family has 2.5 children." Reason: Humans cannot be fractional entities.
- Faulty: "A triangle has sides of length 3 cm, 4 cm, and 8 cm." Reason: The triangle inequality rule is violated; these sides cannot form a triangle.
- Faulty: "A car travels 30 yards per gallon of gas." Reason: Yard is an impractical unit for fuel efficiency.

# 3 Discussion

**Are certain disciplines (e.g., Math, Physics) more challenging for LLMs to process faulty questions correctly?**

Through our experiments, it became evident that questions in Physics posed greater challenges for LLMs compared to those in Mathematics. This may be attributed to the broader range of contextual knowledge required to answer Physics questions correctly. While mathematical problems often rely on clear numerical relationships and rules, Physics questions frequently involve abstract concepts like units, vector properties, and real-world plausibility. For instance, a physics question about velocity being negative is more nuanced than basic arithmetic inconsistencies in math. These subtleties often lead to systematic errors in LLM reasoning.

**What are common fault recognition patterns for LLMs?**

Analyzing LLM responses revealed several recurring patterns where the models struggled to identify faulty questions:

- 
- **Unit/Numerical Misinterpretations:** LLMs often misinterpret units or numerical relationships. For example, questions with unrealistic units such as "30 yards per gallon" are often processed without the model recognizing the impracticality of the unit conversion.
- Unrealistic Real-World Settings: Faulty questions involving implausible scenarios, such as a triangle with sides 3 cm, 4 cm, and 8 cm, are rarely flagged as invalid. This indicates that LLMs struggle to incorporate basic real-world constraints into their reasoning.
- **Misclassification of Similar Concepts:** LLMs frequently confuse closely related terms. For example, velocity is treated as interchangeable with speed, despite the critical distinction that velocity can be negative while speed cannot. This shows a lack of precise contextual understanding of physical terms.
- **Surface-Level Plausibility:** When a faulty question appears superficially valid (e.g., "2.5 children per family"), the model tends to process it

without recognizing the logical inconsistency, relying on pattern matching rather than deeper reasoning.

# 4 Conclusion

Our findings indicate that Physics questions are more challenging for LLMs than Mathematics due to the broader contextual knowledge and abstract reasoning required. Common fault recognition issues include unit and numerical misinterpretations, failure to identify unrealistic real-world scenarios, and confusion between closely related concepts like velocity and speed. Furthermore, the models struggled more with questions of higher complexity, particularly those requiring multi-step logical reasoning.

These results reveal the current limitations of LLMs in handling nuanced and context-dependent questions, underscoring the need for improvements in reasoning capabilities, context retention, and real-world knowledge integration. Future research could expand this analysis across additional scientific disciplines and explore methods to enhance fault recognition in LLMs, contributing to the development of more robust AI systems.