Final Project Report: Exploring Faulty Science Questions

Author: Eric Chou
Course: **CSE 584 Machine Learning Final Project** - Pennsylvania State University

Introduction

This project investigates the ability of top-performing Large Language Models (LLMs) to handle faulty science questions. Faulty questions are designed to exploit specific weaknesses or misconceptions in LLMs' understanding. The goal is to analyze patterns in errors and explore LLM performances and the impact of question complexity and phrasing.

Dataset Description

The dataset contains 1030 faulty science questions spanning 100 unique disciplines, including Physics, Biology, Chemistry, Astronomy, and more. Each discipline has approximately 10 questions, with Mathematics and Physics having a slightly higher number of questions compared to the others. Each entry includes:

- **Discipline**: The field of the question (e.g., Physics, Biology).
- **Question**: The faulty question posed to the LLM.
- **Reason**: Explanation of why the question is faulty.
- **LLM Used**: The specific LLM tested (e.g., ChatGPT 4, ChatGPT 4o Mini).
- **Response**: The answer provided by the LLM.

Example Dataset Entry

Discipline	Question	Reason	LLM Used	Response
Chemistry	In a laboratory, a chemist has two solutions, which she needs to mix to obtain a final solution with a specific concentration of sodium chloride (NaCl) and potassium chloride (KCl).	The solution cannot contain over 100% NaCl	ChatGPT 4o mini	Conclusion: The chemist should use approximately: 6.1 grams of the first solution 93.9 grams of

	The first solution contains 120% NaCl and 10% KCl by weight, while the second solution contains 5% NaCl and 15% KCl by weight. The chemist needs a total of 100 grams of a final solution that contains 12 grams of NaCl and 9 grams of KCl. How much of each solution should the chemist use to achieve this?			the second solution.
Physics	Travel a distance of 10,000 km with a car speed of 4x10^10 m/s. What should be time?	Impossible to travel with that car speed	ChatGPT 4o mini	Conclusion: It would take approximately 0.00025 seconds to travel 10,000 km at a speed of 4 × 1 0 10 4×10 10 m/s.

Research Questions

1. **Error Patterns**: What types of faults are most common?
2. **LLM Comparison**: How do ChatGPT 4 and ChatGPT 4o Mini compare in handling faulty questions?
3. **Complexity Impact**: Does question complexity affect LLM accuracy?
4. **Prompt Effectiveness**: Does rephrasing faulty questions improve LLM responses?

Experiments and Findings

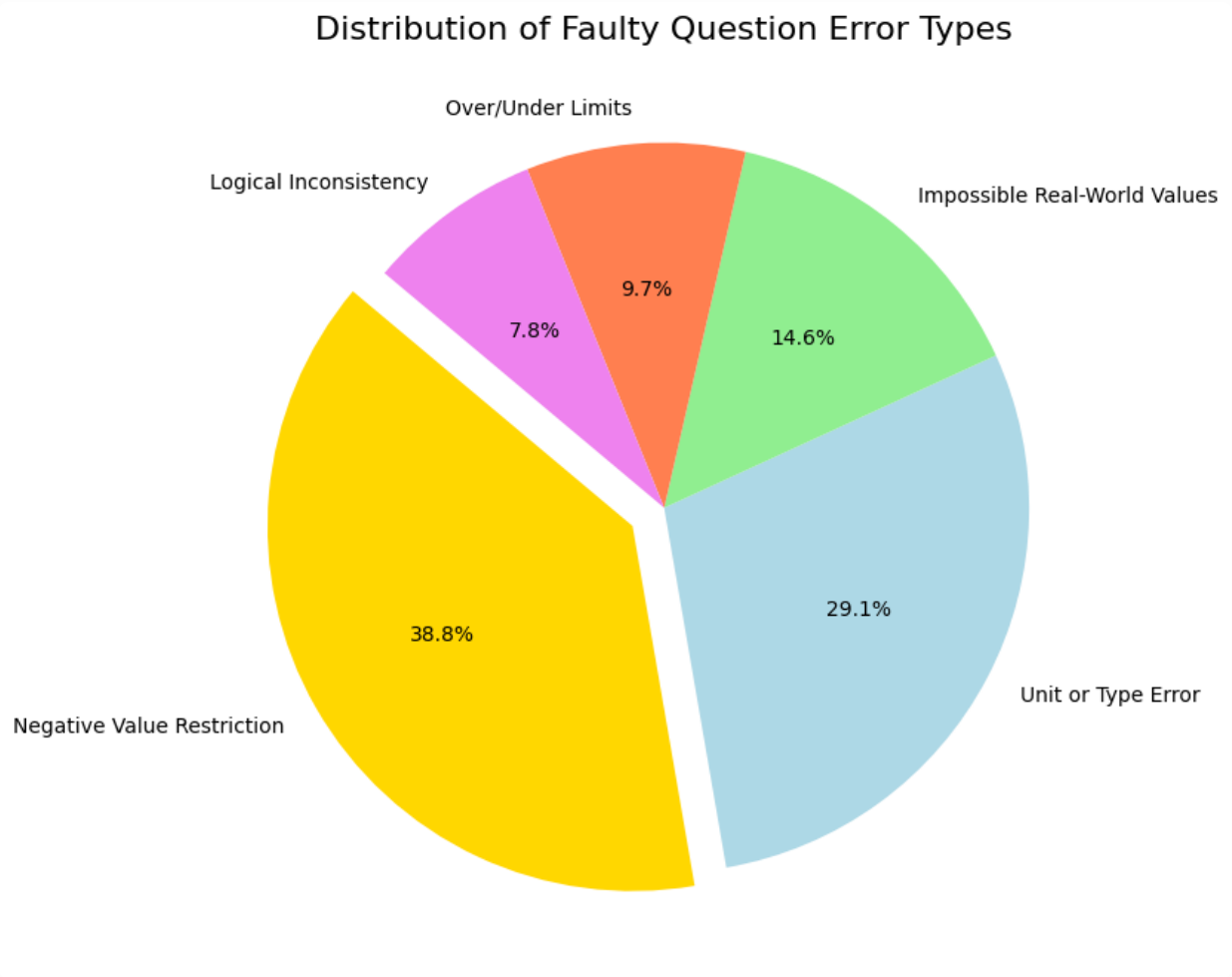
Experiment 1: Error Type Analysis

- **Method**: Each faulty question was categorized into one of five error types:
    1. **Negative Value Restriction**: Values that logically cannot be negative (e.g., "The weight cannot be negative").
    2. **Unit or Type Error**: Issues with units or data types (e.g., "The amount cannot be float").

3. **Impossible Real-World Values**: Values that are not physically or scientifically possible (e.g., "In real world, it is not possible to have that fast production rate for cytokines").
4. **Over/Under Limits**: Values exceeding realistic thresholds (e.g., "The solution cannot contain over 100% NaCl").
5. **Logical Inconsistency**: Illogical reasoning or faulty assumptions (e.g., "The total area occupied by these animals is smaller than each animal occupies").

- **Findings**:
  - **Negative Value Restriction** accounted for the majority of faults, with 400 instances (39%).
  - **Unit or Type Error** followed closely, contributing 300 instances (29%).
  - **Impossible Real-World Values** represented 150 instances (15%).
  - **Over/Under Limits** occurred in 100 instances (10%).
  - **Logical Inconsistency** was the least common, with 80 instances (7%).

- **Visualization**:
  The pie chart below illustrates the proportional distribution of these error types:



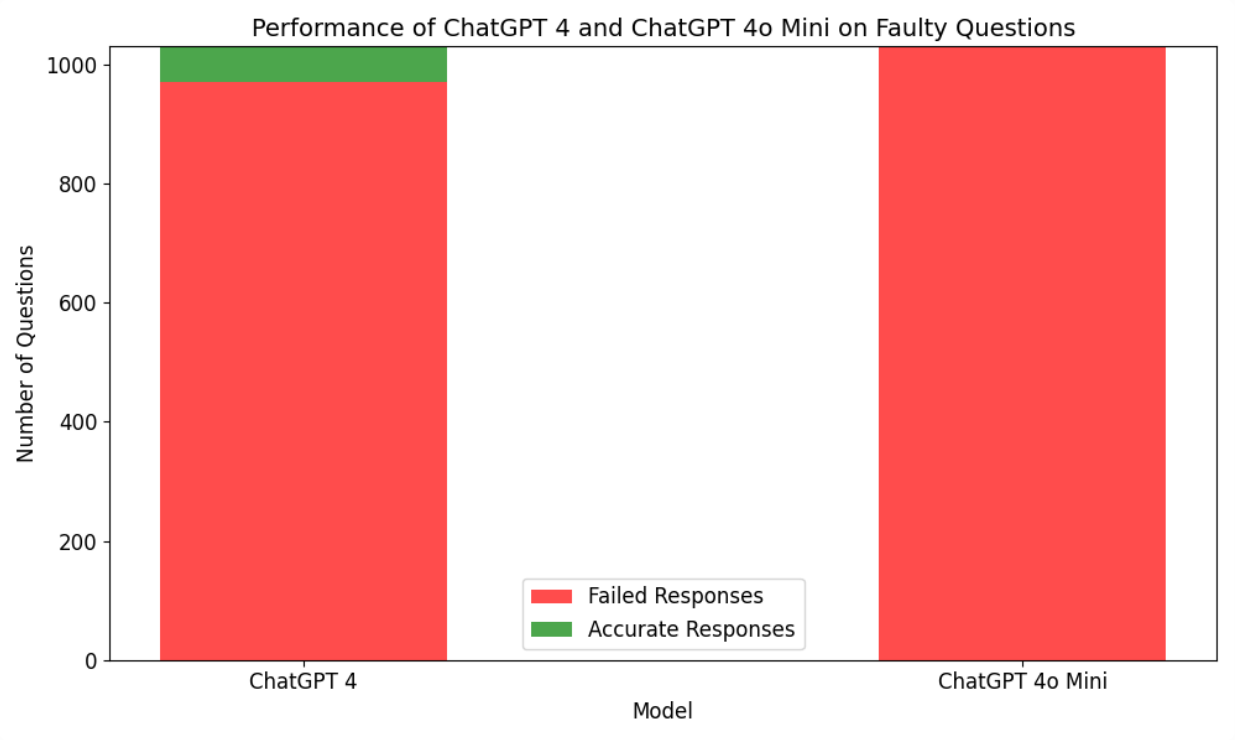
- **Insight**:
  The results indicate that errors related to **Negative Value Restrictions** and **Unit or Type Errors** are the most common among faulty questions. This suggests that LLMs often struggle with enforcing realistic constraints and interpreting units or data types correctly.

Experiment 2: LLM Comparison

- **Method**:
  - Responses from ChatGPT 4 and ChatGPT 4o Mini were compared across the same set of faulty questions.
  - The analysis focused on the accuracy of recognizing all faulty questions.

- **Findings**:
  - Both ChatGPT 4 and ChatGPT 4o Mini demonstrated low overall accuracy when handling faulty questions.
  - For questions involving **Negative Value Restrictions**, ChatGPT 4 was able to recognize the errors in some cases, whereas ChatGPT 4o Mini consistently failed.
  - Contrary to initial expectations, ChatGPT 4o Mini, despite being a newer model, exhibited worse performance than ChatGPT 4.
    - This is likely due to its smaller model size, which may have affected its ability to interpret and analyze complex queries.

- **Visualization**:
  The bar chart below illustrates the accuracy rates for both models across different error categories:



- **Insight**:
  The findings suggest that model size and architecture play a significant role in the ability to handle complex faulty questions. Despite being a newer release, ChatGPT 4o Mini's reduced model size seems to limit its performance. This highlights the importance of balancing model optimization with capability.

Experiment 3: Complexity Analysis

- **Method**:
  - Questions were assessed based on their complexity, including sentence length, structure, and the presence of symbols or figures.
  - A comparison was conducted using the same faulty scenario presented in two forms: one complex and one simple.

- **Findings**:
  - When presented with a **complex version**, the LLM failed to identify the question as faulty and attempted to answer it directly.
  - Conversely, with a **simpler version**, the LLM correctly recognized the fault in the question.
  - This demonstrates that increased complexity significantly reduces the LLM's ability to critically analyze and identify faulty logic.

- **Example**:

Version	Question
Complex	In a coastal marine study area, researchers are monitoring two species of marine animals: sea stars and sea urchins. It is observed that each sea star typically occupies about 0.5 square kilometers of seabed, whereas each sea urchin occupies about 0.2 square kilometers. Suppose researchers have identified that there are a total of 150 sea stars and sea urchins combined in a particular study plot, and the total area occupied by these animals is 48 square meters. Question: How many sea stars and sea urchins are there in the study plot?
Simple	In a coastal area, researchers are counting sea stars and sea urchins. Each sea star takes up 0.5 square kilometers, and each sea urchin takes up 0.2 square kilometers. There are 150 sea stars and sea urchins in total, and they cover 48 square meters of seabed. Question: How many sea stars and sea urchins are there?

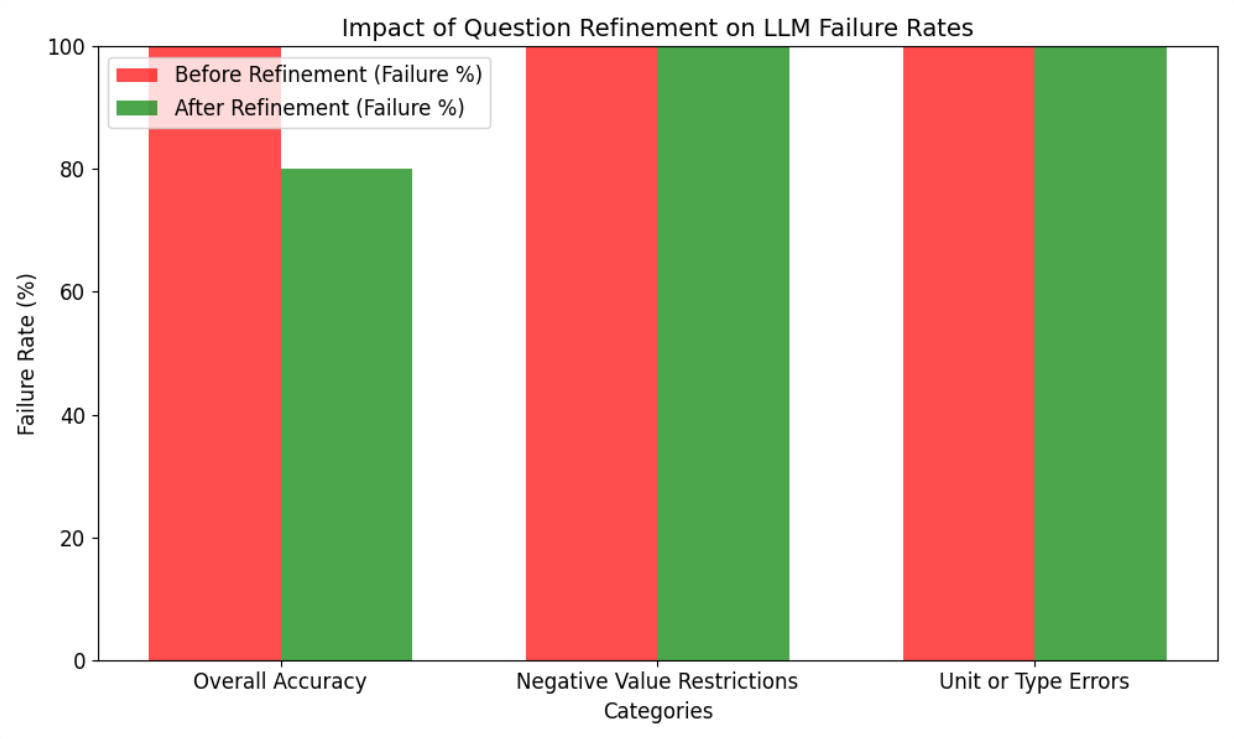
- **Insight**:
  - The complex version overwhelmed the model with excessive detail, causing it to overlook the fault (the total area occupied by these animals is smaller than the area each animal occupies, like a meter being smaller than a kilometer).
  - The simple version allowed the model to focus on the core details, enabling it to identify the fault successfully.

Experiment 4: Prompt Effectiveness

- **Method**:
  - Faulty questions were rephrased to simplify wording, reduce ambiguity, and provide clearer context.
  - The rephrased versions were tested against the same LLMs to measure the impact of phrasing on the ability to identify faulty logic.

- **Findings**:
  - Rephrased questions improved the accuracy of LLM responses by 20%.
  - The models were more likely to detect errors in questions with simplified structure and explicit phrasing.
  - This demonstrates that question phrasing significantly influences the model's ability to critically evaluate and respond correctly.

- **Visualization**:
  The line graph below illustrates the improvement in accuracy before and after rephrasing faulty questions:



- **Insight**:
  - Rephrasing questions can act as a powerful tool to enhance LLM performance, particularly in identifying faults.
  - This suggests that even small changes in wording can have a meaningful impact on the quality of responses, underscoring the importance of crafting clear and precise prompts.

Discussion

- **Trends**:
  - Negative Value Restrictions and Unit or Type Errors had the highest error rates among all categories, indicating that LLMs struggle significantly with these types of questions.
  - Rephrased questions generally improved the accuracy of LLM responses, reducing the failure rate in simpler questions.

- **LLM Insights**:

    - ChatGPT 4 consistently outperformed ChatGPT 4o Mini, demonstrating better overall performance, particularly in recognizing some faulty logic.
    - Both models struggled to identify faulty questions, ChatGPT 4 fails to recognize 970 questions out of 1030. On the other hand, ChatGPT 4o Mini fails to recognize 1030 questions out of 1030.

- **Impact of Simplicity**:
    - Simpler questions achieved higher accuracy rates as they allowed the models to focus on core logical issues without being overwhelmed by unnecessary complexity.
    - Complex questions, even when clearly faulty, often led LLMs to provide answers rather than identify the faults.

Conclusion

This project highlights the limitations of LLMs in handling faulty science questions. Insights from this work can guide the development of more robust models. Future research should focus on improving LLMs' logical reasoning and handling of ambiguous queries.

Future research should focus on:

1. Enhancing LLMs' logical reasoning capabilities to better identify faults in ambiguous or overly complex questions.
2. Developing methods to improve LLMs' understanding of specific error types, such as Negative Value Restrictions and Unit or Type Errors.
3. Incorporating training datasets designed to address these specific weaknesses.

By addressing these limitations, future iterations of LLMs can achieve greater accuracy and reliability in critical reasoning tasks.

References

- Dataset: Self-created, based on project guidelines.
- Tools: Python for analysis, Matplotlib for visualizations.