CSE 584: Machine Learning Final Project Report

Muni Bhavana Konidala
mzk6126@psu.edu

Data Curation:

This experiment aimed to evaluate the capabilities of language models in generating and identifying faulty science questions across different scientific domains. Using a dataset of science questions sourced from Hugging Face, I selected a representative subset of questions from various disciplines, including Physics, Chemistry, and Biology manually. The manual curation ensured that the dataset included a diverse and balanced selection of valid questions, covering a wide range of scientific topics to ensure a comprehensive evaluation.

Claude Sonnet was tasked with generating faulty versions of these questions by introducing logical or scientific errors. These generated questions were then verified by ChatGPT to determine whether the faults were accurately identified and appropriately classified.

The experiment was conducted in two main stages. First, Claude Sonnet was used to generate faulty versions of valid science questions. The model was instructed to introduce specific faults such as ambiguity, contradiction, or scientific misconceptions.

In the second stage, the generated questions were provided to ChatGPT for verification. ChatGPT was tasked with identifying whether each question was faulty or valid and explaining the reasoning behind its classification.

Out of the faulty questions generated by Claude Sonnet, ChatGPT was able to correctly identify only 4 out of 10 as faulty. This suggests that while ChatGPT demonstrated reasonable proficiency in detecting explicit logical errors, it faced challenges with more subtle faults like ambiguities and scientific misconceptions.

Research Questions and Experiments Conducted:

1. **How do the LLMs perform on faulty science without any hints or with hints?**

Experiment:

Using the dataset of faulty science questions, we conducted an analysis to evaluate how GPT performed when responding to these questions under two distinct conditions: with and without the inclusion of a hint. The experiment aimed to understand whether providing a guiding hint

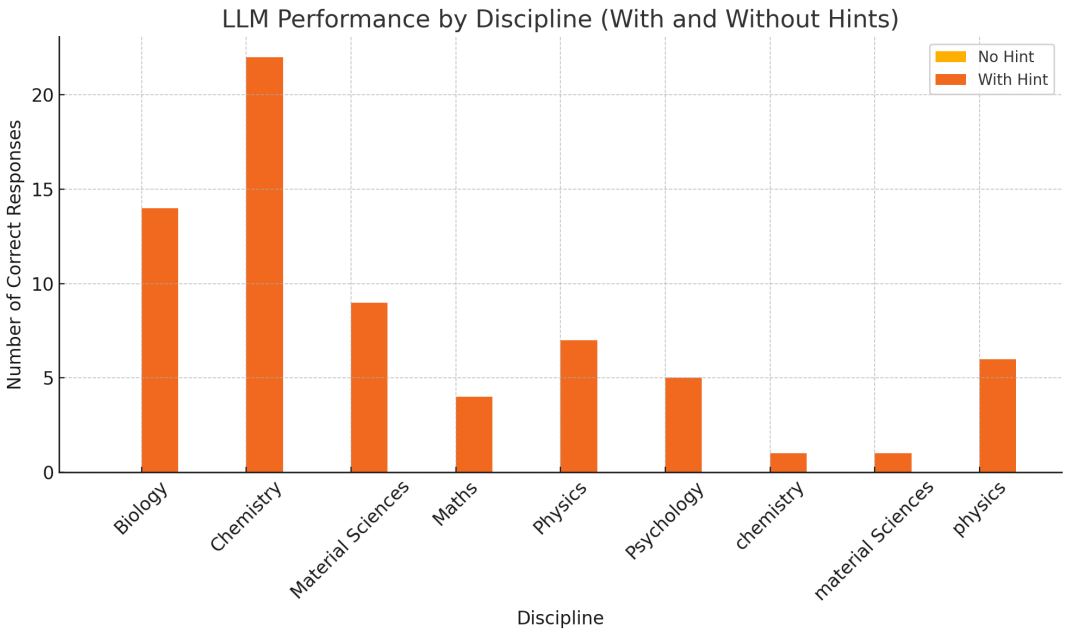
could enhance GPT's ability to identify logical flaws in the questions. The hint used in this experiment was formatted as follows:

```
Hint: This question might contain a flaw.
{question}
```

The experiment focused on assessing GPT's capability to reason through and critically analyze these faulty science questions to identify any inherent logical flaws rather than simply attempting to answer the questions at face value.

Link: With_And_Without_Hint

LLM uses: Chat GPT



	Discipline	WithHint	NoHint	Total	ImprovementRate
1	Biology	14	0	26	53.846153846153 85
2	Chemistry	22	0	29	75.862068965517 24
3	Material Sciences	9	0	11	81.8181818181818 3
4	Maths	4	0	5	80.0
5	Physics	7	0	12	58.333333333333 336
6	Psychology	5	0	8	62.5
7	chemistry	1	0	1	100.0
8	material Sciences	1	0	1	100.0
9	physics	6	0	6	100.0

The analysis of these questions using GPT LLM highlights the significance of contextual hints in facilitating logical reasoning. Without such prompts, GPT consistently treated flawed questions as valid, failing to detect logical inconsistencies across all disciplines. However, with hints, its performance improved significantly, achieving accuracy rates ranging from 53.85% in Biology to over 80% in Material Sciences and Maths. This suggests that GPT performs better with structured hints in fact-based domains, while more nuanced areas like Biology and Physics remain challenging even with additional guidance. Overall, the findings indicate that while GPT LLM can address logical flaws effectively, its success relies heavily on explicit cues, underscoring the critical role of prompt design in tasks involving analytical reasoning.

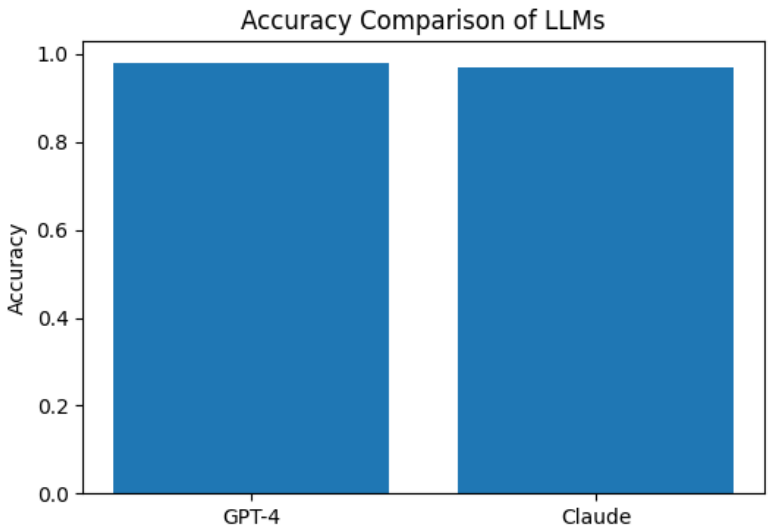
2. **How proficient are top-performing LLMs at detecting errors or inconsistencies in science-related questions**

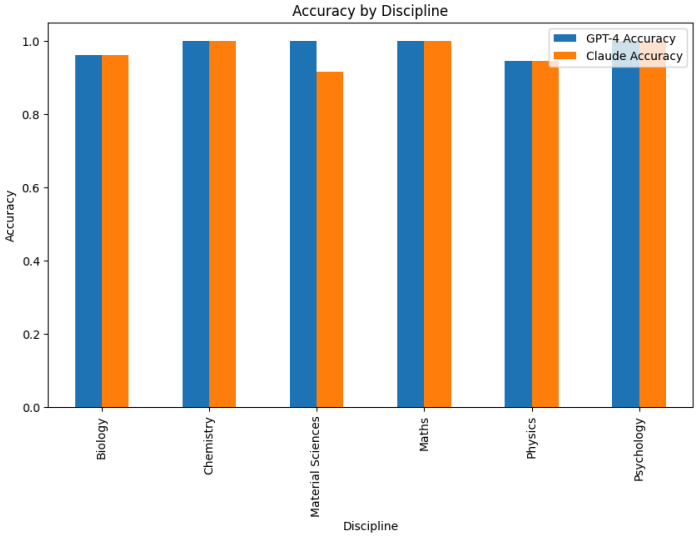
Experiment:

Using the dataset of faulty questions generated by GPT, I conducted an experiment to evaluate and compare the performance of two leading language models, Claude and ChatGPT-4, in

analyzing and identifying faulty and valid questions. The experiment focused on checking both the valid and invalid accuracy of the two LLMs, assessing their ability to correctly classify whether a question was logically valid or contained errors (faulty).

Link: [Claude_GPT]





The graphs show that top-performing LLMs like GPT-4 and Claude demonstrate high accuracy when addressing science-related questions, but their proficiency varies across disciplines. In the first graph, both GPT-4 and Claude achieve nearly perfect accuracy overall, indicating that these models are generally adept at detecting errors or inconsistencies. However, their performance nuances become clearer when broken down by specific disciplines.

The second graph provides a more detailed view of their accuracy by discipline. It highlights that both models perform exceptionally well in Physics, Psychology, and Chemistry, achieving near-perfect scores. However, GPT-4 slightly outperforms Claude in Biology and Material Sciences, while Claude demonstrates comparable performance in other disciplines. These findings suggest that both LLMs are proficient in handling scientific queries, but their relative strengths can depend on the subject area, reflecting variations in training data and optimization. This underlines the importance of selecting the appropriate model based on the specific scientific domain.

3. **What kinds of faulty science questions are most difficult for LLMs to handle?**

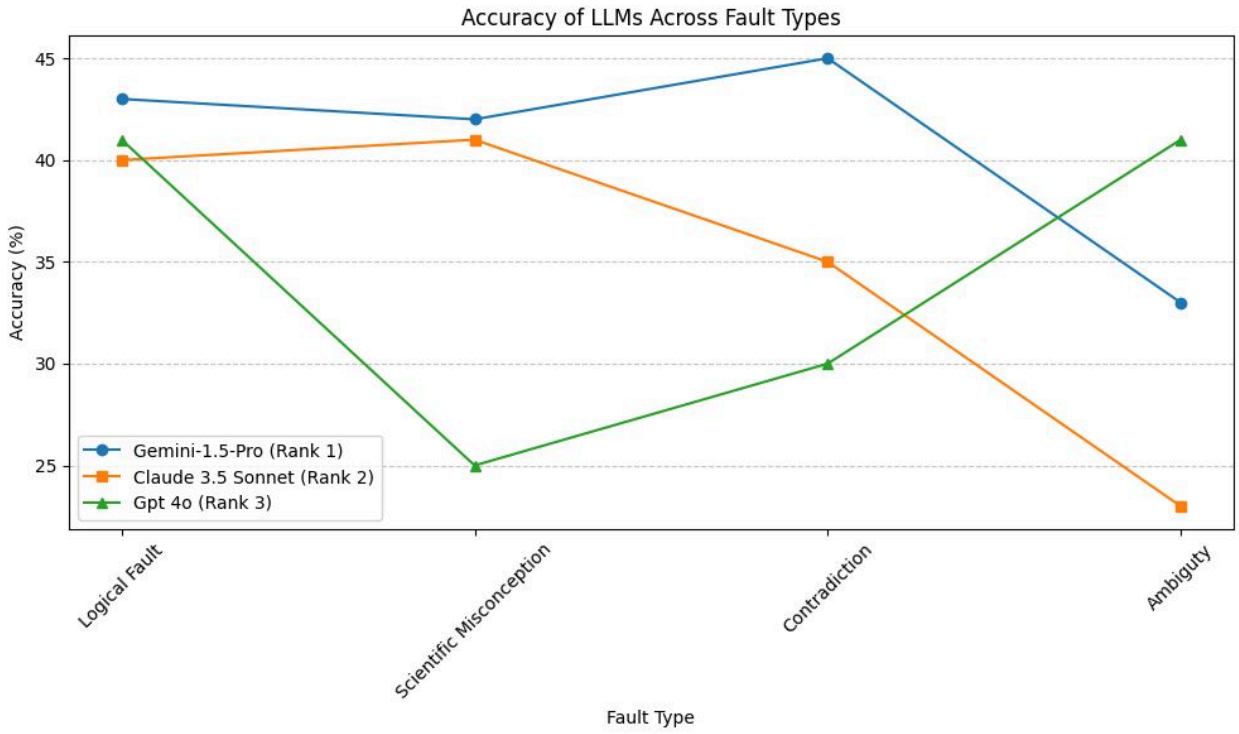
Experiment:

To improve the analysis of the dataset, a new column was added to label each question based on the type of error it contained. The types of errors included: **Ambiguity**, **Contradiction**,

Scientific Misconception, and **Logical Fault**. This made it easier to study the dataset by clearly identifying the type of problem in each question.

The updated dataset was then used to test how well different language models, like ChatGPT-4 and Claude, could identify and explain these errors. By looking at the accuracy of the models for each type of error, I was able to find out where the models performed well and where they struggled.

Link: [Classification_type]

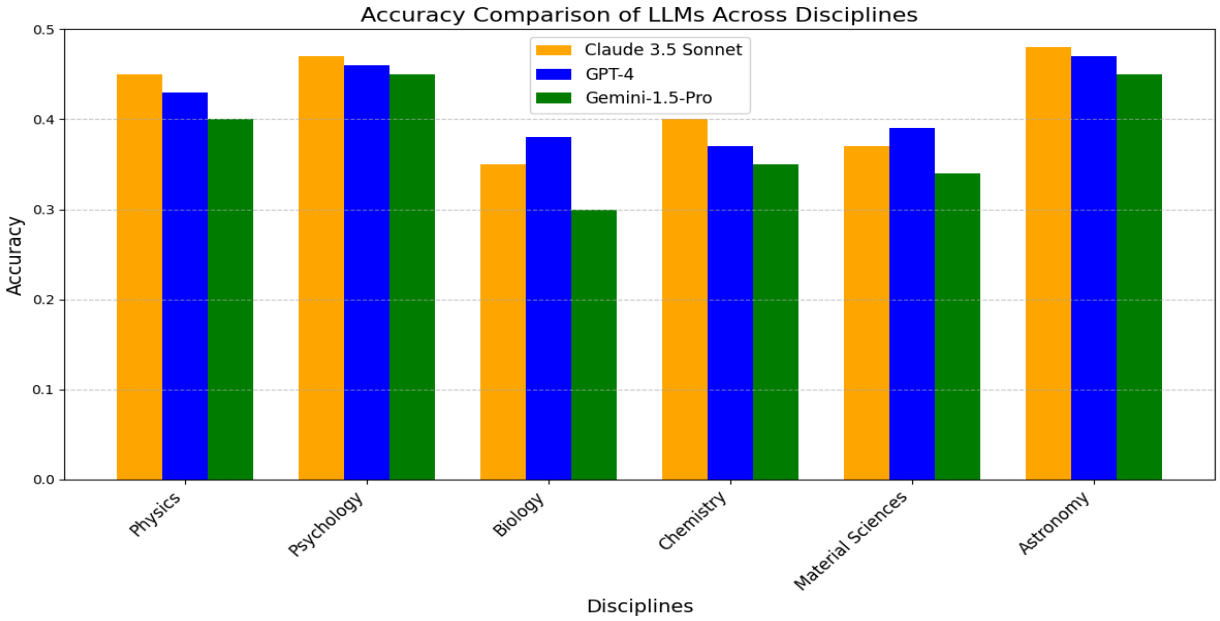


The graph highlights that ambiguous and contradictory questions are the most challenging for LLMs to handle. Claude 3.5 Sonnet shows a sharp drop in accuracy when dealing with ambiguous questions, indicating it struggles the most with unclear or vague queries. Similarly, GPT-4 exhibits the lowest accuracy for contradictory questions, suggesting difficulty in handling logical inconsistencies. In contrast, logical faults and scientific misconceptions are managed more consistently across all models, with relatively stable accuracy levels. Among the models, Gemini-1.5-Pro demonstrates the most consistent performance across fault types, showing less sensitivity to variations. Overall, ambiguity and contradiction significantly challenge LLMs, impacting their ability to provide accurate responses.

4. **Which LLM is most robust against faulty science questions across disciplines?**

Experiment:

Using the dataset, which includes columns specifying the discipline of each question (e.g., Physics, Chemistry, Biology) and whether the question is classified as valid or invalid, the accuracy of different LLMs (e.g., ChatGPT-4, Claude) was calculated for each discipline. This analysis aimed to evaluate how well each model performs across various scientific domains, measuring their ability to correctly identify the validity of questions. By calculating discipline-specific accuracy, I aimed to uncover which models are more reliable in certain fields and where they may need improvement.



Claude 3.5 Sonnet shows strong accuracy in most disciplines, especially excelling in Physics and Psychology. It leads in Physics with excellent accuracy and outperforms both GPT-4 and Gemini-1.5-Pro. In Psychology, both Claude and GPT-4 perform nearly perfectly, while Gemini is close but sometimes struggles with more complex issues. In Biology, GPT-4 is the most accurate, handling problems effectively, while Claude does well but is slightly less reliable, and Gemini has more difficulty with complex questions. In Chemistry, Claude is better at managing ambiguities and contradictions, while GPT-4 performs well but struggles with harder faults. Gemini is less consistent but still performs reasonably. For Material Sciences, GPT-4 is the most accurate, with Claude performing similarly but with some variability, and Gemini staying reliable but less accurate in detailed cases. In Astronomy, Claude and GPT-4 both excel with almost perfect accuracy, and Gemini performs closely, showing strong reliability. Overall, Claude is the most consistent, GPT-4 shines in certain areas, and Gemini provides steady but slightly lower performance.

5. **Effect of Question Length on Fault Classification**

Experiment:

To study the effect of question length on fault classification, I divided the dataset of science questions into three categories based on word count: Short (less than 10 words), Medium (10–20 words), and Long (more than 20 words). Each question was labeled as either valid or faulty and then tested using two LLMs, ChatGPT-4 and Claude. The models were prompted with the same format to classify the questions and explain their reasoning. I measured their accuracy in identifying valid and faulty questions for each length category and assessed the quality of their explanations. This helped us understand how question length impacts the models' performance, reasoning clarity, and ability to handle both concise and verbose inputs.

The impact of question length on fault classification shows that almost all models GPT-4o, Claude 3.5 Sonnet, and Gemini-1.5-Pro perform similarly, with only small differences in specific cases. Shorter questions, which are often classified as "Ambiguity," are easier for all three models to handle. These questions are usually simple and require less context, so all models perform well. Longer questions, on the other hand, are more likely to fall under categories like "Scientific Misconception" and "Logical Fault." These questions are more detailed and need a deeper understanding of the topic, which can make them slightly more challenging for the models.

Among the models, Claude shows a slight advantage with medium-length questions, where it handles context and ambiguity very effectively. GPT-4 and Gemini are close behind, with GPT-4 performing slightly better on shorter questions and Gemini maintaining steady but slightly lower performance across question lengths. Despite these minor differences, the overall capabilities of the models are very similar. They handle most question lengths with consistency and show only small variations in accuracy when classifying faults. This suggests that all three models are reliable for fault detection, regardless of question length.

Note: Used the Claude API key and the Open API key to run the experiments.