

Name: **Rohan Prasad**
Email: rpp5524@psu.edu
PSU ID: **980707395**

Date: 12/6/2024

CSE 584 Final Project Report

Index

CSE 584 Final Project Report	1
Problem Statement	2
Dataset	2
Exploratory Data Analysis	2
Interesting Research Questions	4
Question 1	4
Question 2	5
Question 3	6
Question 4	6
Experiments	7
Experiment 1: LLM Response Analysis to Flawed Questions	7
Experiment 2: Impact of Domain-Specific Questions	9
Experiment 3: Real-Time Correction Test	11
Experiment 4: Human-in-the-loop assistance or LLM as a Judge	13
References	13

Problem Statement

Collect or create a set of faulty science questions that can fool a top-performing LLM (e.g., ChatGPT, GPT4, Gemini-1.5-Pro, or Claude-3-Opus, etc.)

Dataset

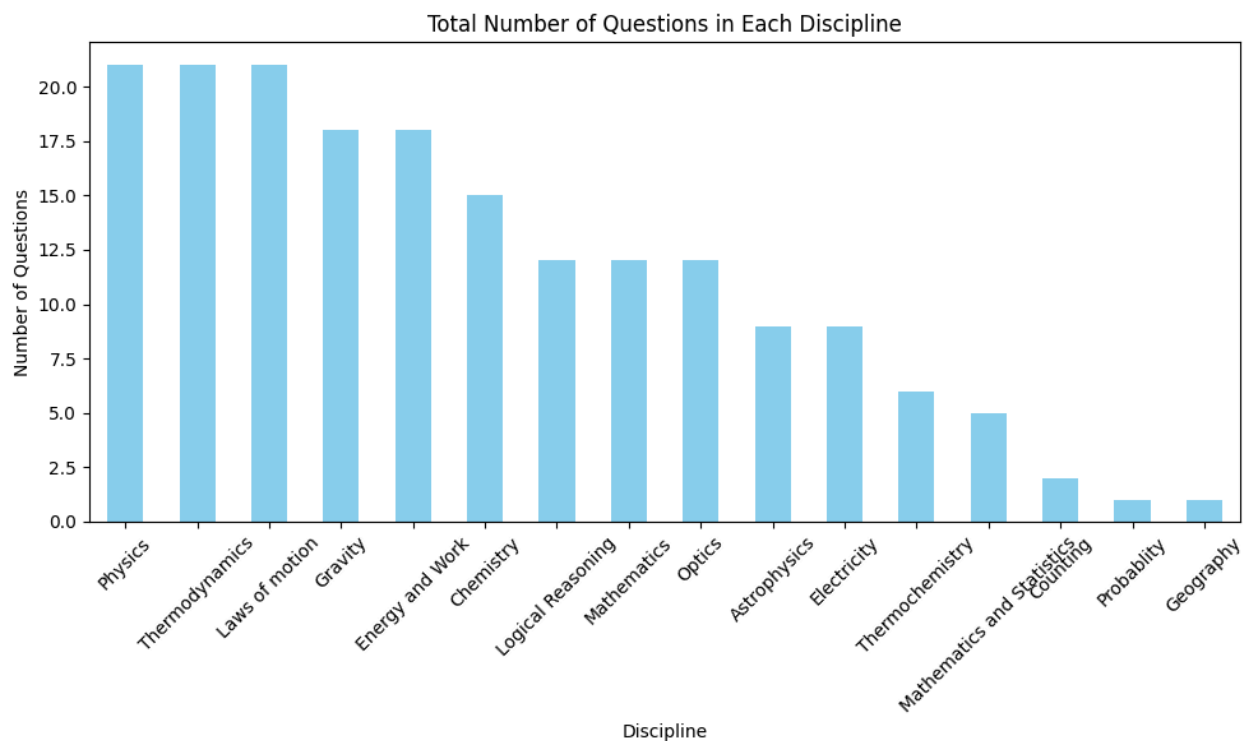
The dataset is in the git repository:

https://github.com/rpp5524/CSE_584_rpp5524/tree/main/Final_Project/dataset

The LLMs chosen for this project are:

- GPT-3
- GPT-4
- LLama 3.2
- Mistral

Exploratory Data Analysis

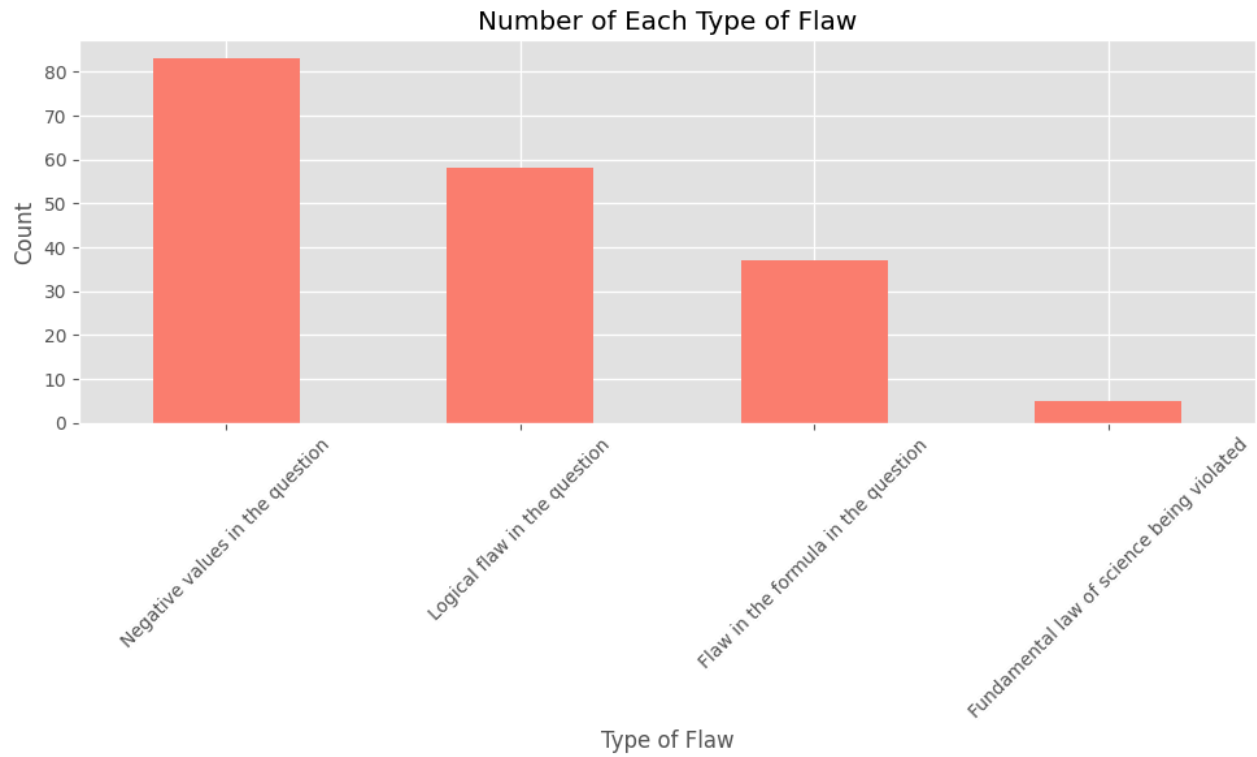


After performing an initial exploratory data analysis, it can be observed that the science questions involved have a variety of disciplines. The disciplines are not perfectly balanced, but there is a slight skew of questions towards Physics, Thermodynamics and Laws of motion disciplines.

The **Type of Flaw** in the question has also been classified:

- Negative values in the question'- The question contains a negative value that is illogical or inappropriate or is not possible to exist.

- 'Fundamental law of science being violated'- The question or response contradicts well-established scientific principles.
- 'Logical flaw in the question'- The question contains logical inconsistencies or contradictions.
- 'Flaw in the formula in the question'- The question includes an incorrect or flawed formula.



Interesting Research Questions

The experiments I have chosen explore how Large Language Models (LLMs) like GPT, LLama and Mistral handle faulty science questions with inherent flaws. It is intriguing and can yield insights into both AI understanding and its potential improvement.

Question 1

How do LLMs respond to scientifically flawed questions?

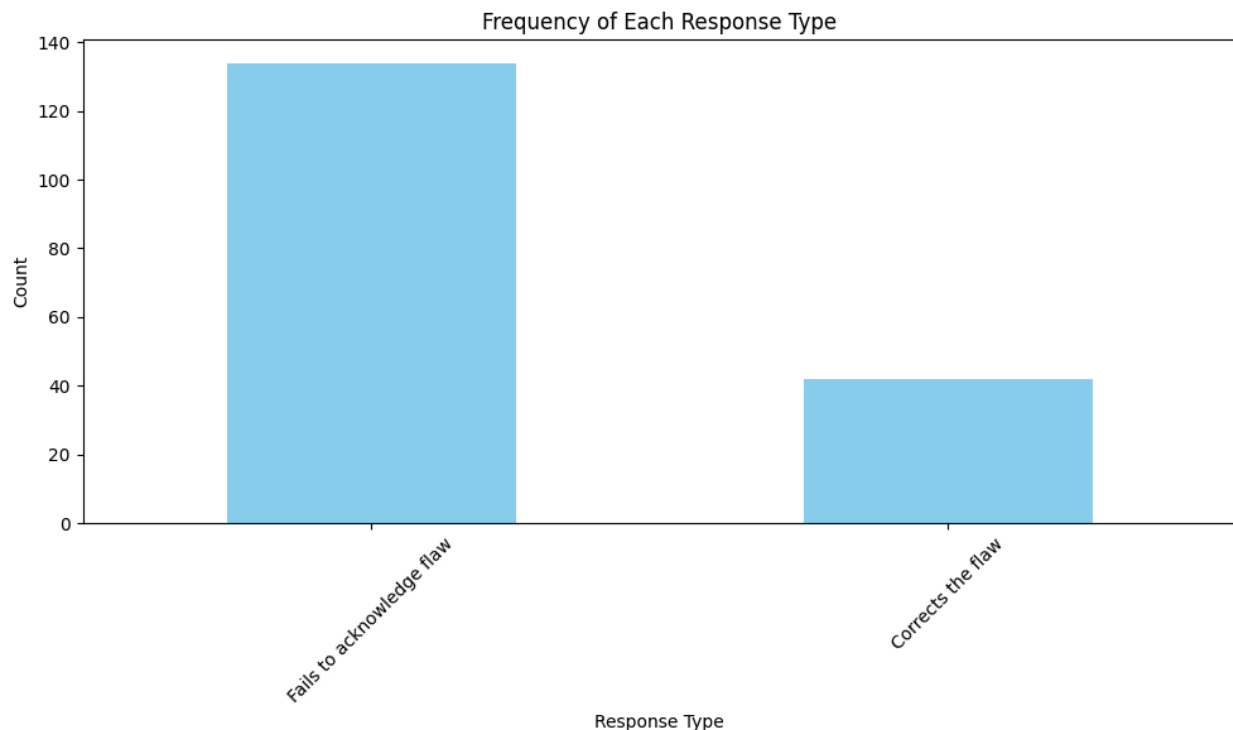
- Do they perpetuate the flaws, correct them, or fail to recognize them altogether?

Significance of the Question

This query assesses how resilient and resistant LLMs are to inaccurate or illogical inputs. Models frequently come into data in real-world applications that could be inaccurate or misleading. Deploying LLMs in critical environments where decision accuracy is vital requires an understanding of how they handle such scenarios.

Observations

It can be observed that the LLMs fail to even acknowledge the flaw or incorrectness in the question. Other times the LLM acknowledges the question but does not even correct it. It proceeds to answer the question.



Question 2

What types of scientific inaccuracies are most likely to mislead LLMs?

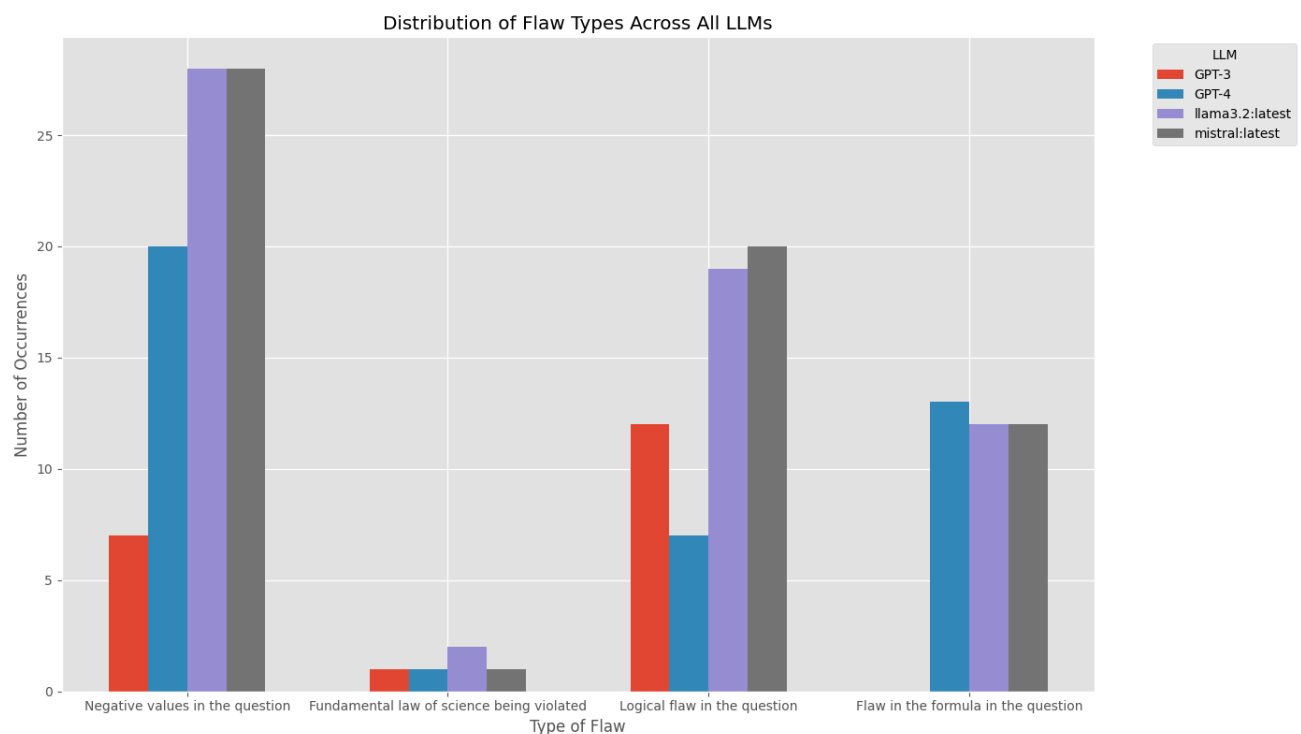
- Are some scientific domains or specific types of flaws (e.g., negative values where none should exist, violations of conservation laws, or violation of fundamental laws of Physics) more likely to result in incorrect responses from LLMs? Or rather what areas or subjects in science can cause LLMs to answer a scientifically incorrect question.

Significance of the Question

Researchers and developers can improve model training by focusing on the precise defect kinds that commonly cause LLMs to be misled. For instance, if problems containing negative absolute values (such as temperature below absolute zero) frequently trick LLMs, more targeted training can be added to improve their handling of these situations. For applications like automated fact-checking, this aids in determining how sensitive LLMs are to different kinds of scientific errors.

Observations

It can be seen that the question has a diverse distribution of the different types of flaws that are in the science question. Distributed evenly across all LLMs.



Question 3

Can LLMs generate explanations that identify and correct errors in the setup of a flawed science question? Will parameter efficient fine tuning the LLM impacts its ability to detect flaws and correct and answer them?

- To what extent can LLMs not only provide correct answers but also educate about why a question is flawed?

Significance of the Question

For instructional reasons, LLMs' capacity to identify mistakes and offer concise justifications and remedies is essential. Because of these characteristics, LLMs can be helpful teaching assistants or tutoring programs that offer answers along with reasoning and helpful critique. Asking LLMs to explain why a question is wrong tests their deeper comprehension and reasoning abilities. This extends beyond simple pattern recognition to more intricate interpretation of concepts and data, which is crucial for advanced AI applications.

Question 4

Can human-in-the-loop systems improve LLMs' detection and correction of flaws in real-time?

Significance of the Question:

LLMs can be complemented by human oversight in high-stakes applications. Human feedback can be given at each step of the LLM generation to help the LLM.

Experiments

Experiment 1: LLM Response Analysis to Flawed Questions

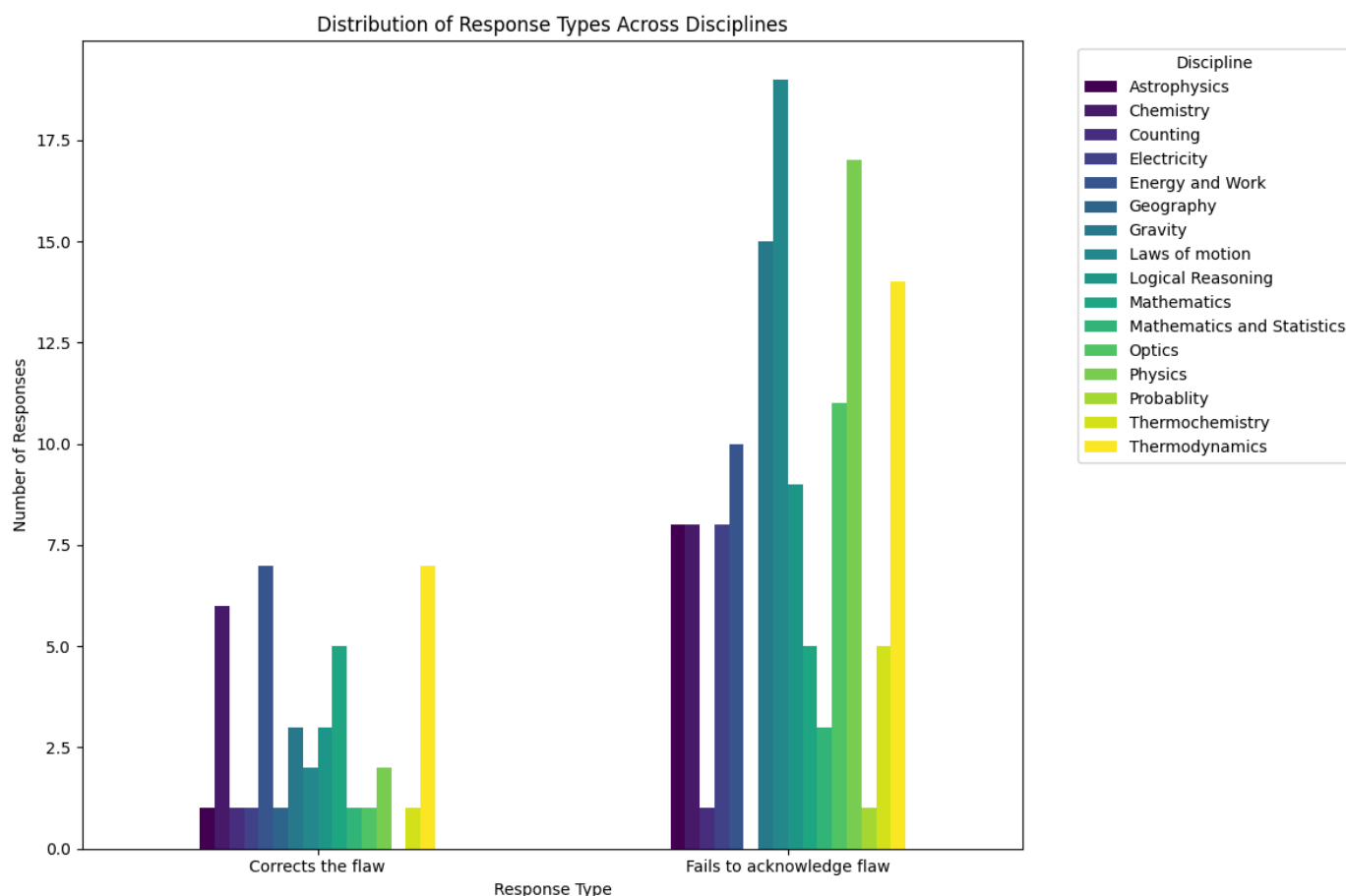
Objective: Assess how LLMs respond to a range of scientifically flawed questions across different domains.

Method:

- I collected faulty science questions that include a variety of flaw types (e.g., impossible values, incorrect scientific laws, logical fallacies, incorrect flaws).

For this I write a script to use **LLM as a Judge** to label the type of flaw and whether the LLM response corrects or acknowledges it.

- I feed these questions to a variety of LLMs and then analyze their responses.
- I categorize the responses into those that correct the flaw in the question, and those that fail to address it.



Question

How do LLMs respond to scientifically flawed questions?

Answer:

The majority of responses seem to fall under the category "Fails to acknowledge flaw." This suggests that LLMs often provide answers without identifying that the question is flawed.

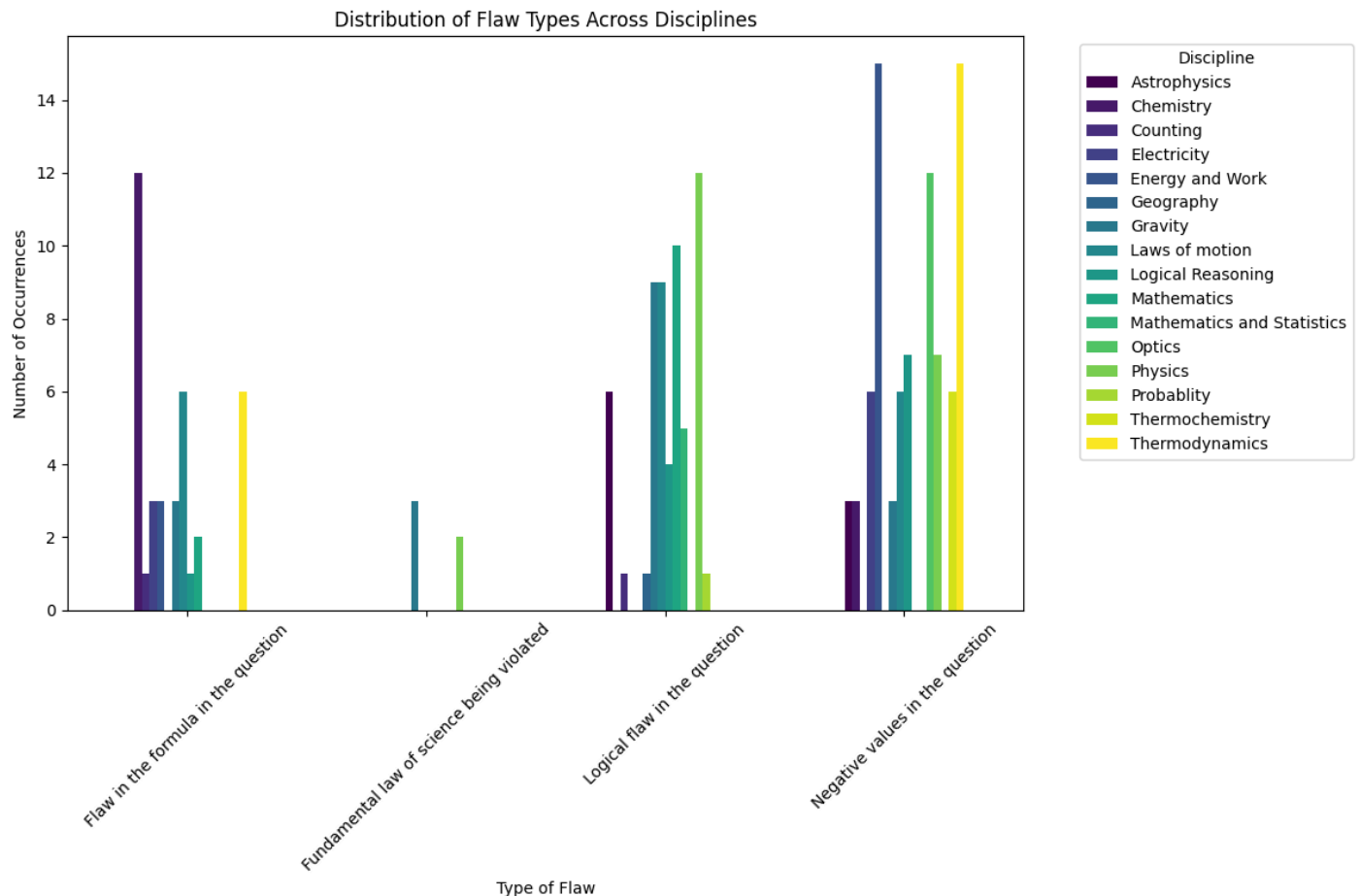
The bars for "Corrects the flaw" are present but significantly smaller in magnitude compared to "Fails to acknowledge flaw." This indicates that while LLMs can identify and correct flaws, they do so much less frequently.

Some disciplines show a higher likelihood of flaws being corrected (e.g., certain fields like "Mathematics" or "Physics"). This may indicate that LLMs perform better in fields where training data is richer or the concepts are more structured. Disciplines like "Logical Reasoning" and "Thermodynamics" seem to have a high count of unacknowledged flaws. This suggests that LLMs might struggle more with reasoning-intensive or specialized scientific areas.

Experiment 2: Impact of Domain-Specific Questions

Objective: To find out what types of domains impact the ability of the LLM to answer scientifically flawed questions.

Method: I grouped the disciplines with respect by the type of flaw in the question and interesting results can be observed



Question: **What types of scientific inaccuracies are most likely to mislead LLMs?**

Answer:

Logical Flaws in the Question seems to have a significant presence across multiple disciplines. They often confuse LLMs because they are not explicitly rule-based reasoning engines, which may not prepare them well for resolving such inconsistencies.

Negative Values also appear frequently across several disciplines, which can easily mislead LLMs if the concept of negative values in a given context (e.g., energy, resistance) isn't explicitly reinforced in training data. Disciplines like **Electricity, Mechanics and laws of motion, Optics** suffer from this.

Incorrect Formulas in the question is concentrated in specific disciplines, such as **Astrophysics, Mathematics**, where formulas are more structured and critical for problem-solving. LLMs struggle here because of partial recall from training data. Misleading formulas are a result of memorization rather than true mathematical reasoning.

Physics, Thermodynamics, and Chemistry frequently appear for flaws like "Negative Values in the Question" and "Logical Flaws." This is because these subjects rely heavily on precise numerical and logical constraints (e.g., absolute zero in thermodynamics, conservation laws in physics).

Violations of fundamental laws appear in domains like **thermodynamics, Gas laws, Laws of motion, Energy and work, gravitation**(e.g., conservation laws, principles) appear less frequently This may be because LLMs are generally better at handling well-documented, widely-known scientific laws that are prominent in training datasets.

Some potential reasons why LLMs behave like this is because they are trained on general-purpose datasets. For example, while common knowledge about formulas is well-documented, **nuanced edge cases** may not be adequately represented.

Disciplines like **Physics, Thermodynamics and Gravitation** are inherently rule-based and structured. Misleading flaws in formulas or logical inconsistencies are harder for them to process.

Experiment 3: Real-Time Correction Test

Objective: Here I determine if training LLMs on high-quality, domain-specific datasets reduces the likelihood of them being fooled by flawed questions. Whether LLMs can detect and correct scientific flaws in questions. I also identified the impact of Parameter-Efficient Fine-Tuning (PEFT) (using LoRA) on the model's ability to improve responses to flawed questions.

Method:

Original Model: **Phi2 Microsoft Model.**

Fine-Tuning Technique: **LoRA for parameter-efficient fine-tuning.**

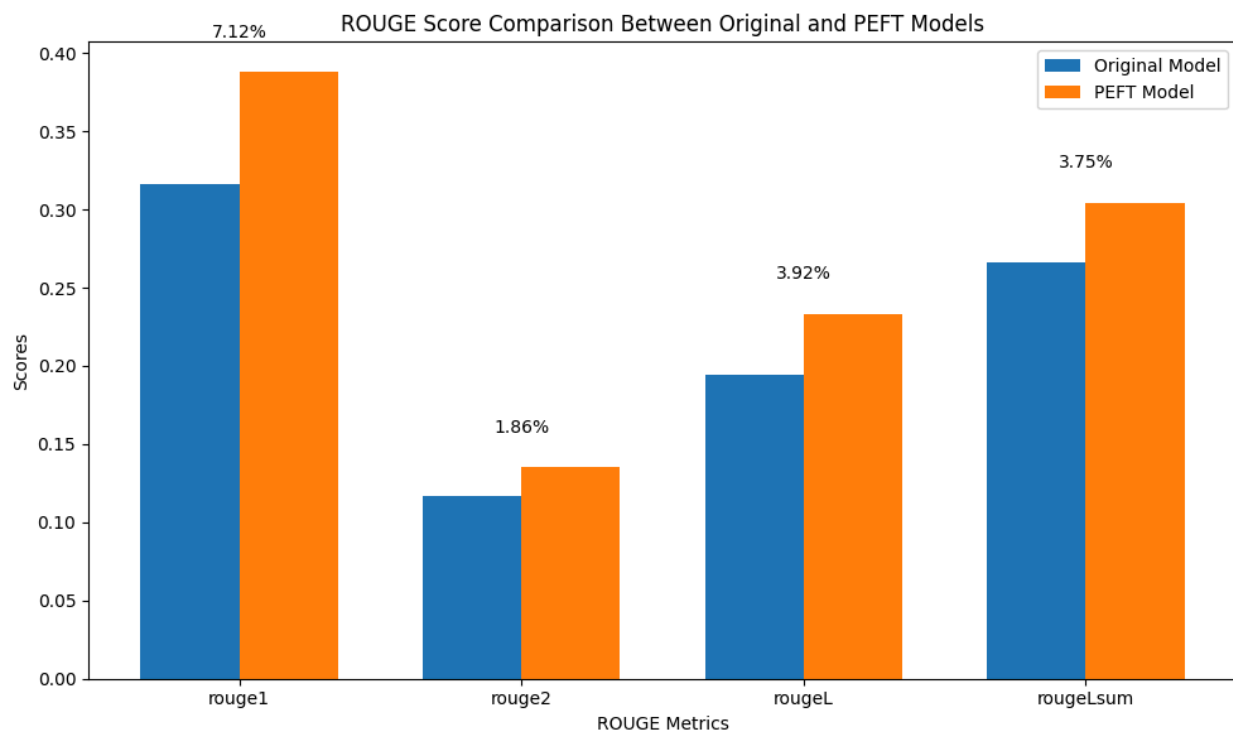
Notebook with Experiments:

https://colab.research.google.com/drive/111lu0g38TP_dYOevfk2vtAkCHuGZIxNN?usp=sharing

I set up two LLMs – one a general base model, another with enhanced training on scientifically accurate content with questions with scientific inconsistencies in them.

Testing: I generated a dataset similar to the one earlier but this dataset has science facts and formulas and general rules and laws that science need to follow. This contains SQA Questions that are factually correct with factually current answers

Metrics: I then measure the ROUGE scores of the base model



ROUGE Scores: Base model vs PEFT model

Question: ***Can LLMs generate explanations that identify and correct errors in the setup of a flawed science question? Will tuning the LLM impacts its ability to detect flaws and correct and answer them?***

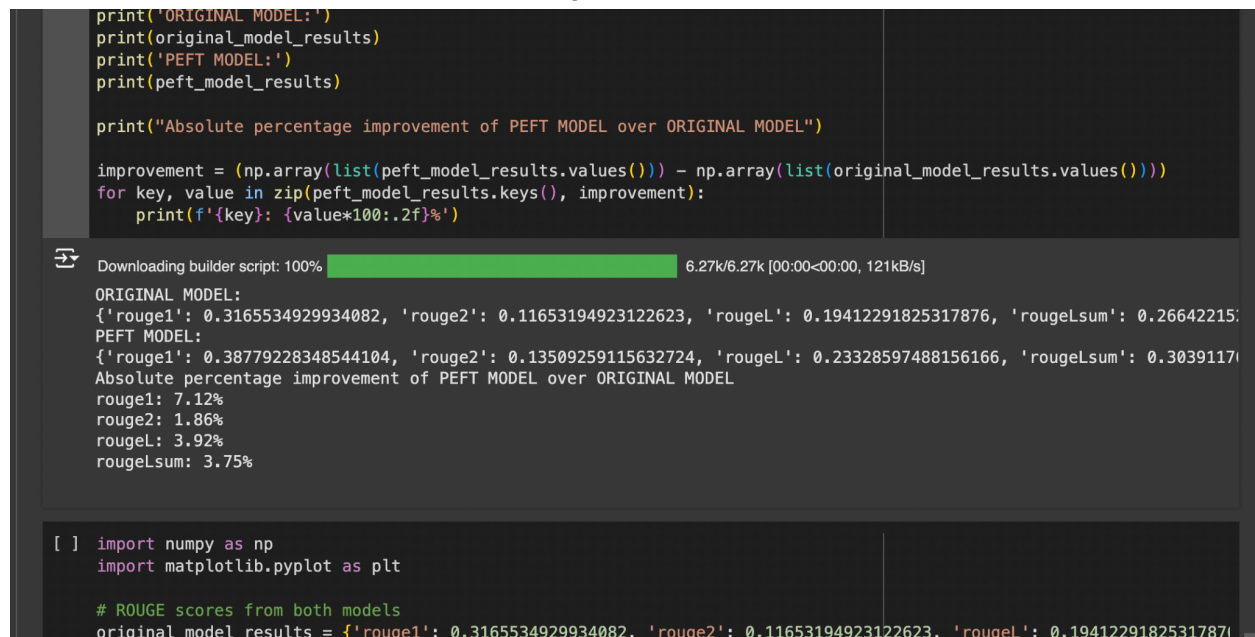
Answer:

Yes, fine-tuned LLMs can effectively detect and correct errors in flawed questions. The fine-tuned Phi2 model demonstrated significant improvements in detecting flaws and generating scientifically sound corrections.

The fine-tuned model showed absolute ROUGE improvements of 7.12% (ROUGE-1) and 3.92% (ROUGE-L), indicating better detection of relevance, coherence, and fluency.

The ability to generate explanations that explicitly identify flaws and correct them improved significantly after fine-tuning.

Screenshot of Parameter efficient fine tuning



```
print('ORIGINAL MODEL:')
print(original_model_results)
print('PEFT MODEL:')
print(peft_model_results)

print("Absolute percentage improvement of PEFT MODEL over ORIGINAL MODEL")

improvement = (np.array(list(peft_model_results.values())) - np.array(list(original_model_results.values())))
for key, value in zip(peft_model_results.keys(), improvement):
    print(f'{key}: {value*100:.2f}%')
```

Downloading builder script: 100% 6.27k/6.27k [00:00<00:00, 121kB/s]

ORIGINAL MODEL:
{'rouge1': 0.3165534929934082, 'rouge2': 0.11653194923122623, 'rougeL': 0.19412291825317876, 'rougeLsum': 0.26642215}

PEFT MODEL:
{'rouge1': 0.38779228348544104, 'rouge2': 0.13509259115632724, 'rougeL': 0.23328597488156166, 'rougeLsum': 0.3039117}

Absolute percentage improvement of PEFT MODEL over ORIGINAL MODEL
rouge1: 7.12%
rouge2: 1.86%
rougeL: 3.92%
rougeLsum: 3.75%

```
[ ] import numpy as np
import matplotlib.pyplot as plt

# ROUGE scores from both models
original_model_results = {'rouge1': 0.3165534929934082, 'rouge2': 0.11653194923122623, 'rougeL': 0.19412291825317876}
```

Experiment 4: Human-in-the-loop assistance or LLM as a Judge

Objective: Here I determine if human-in-the-loop feedback or **LLM as a judge** assist the LLM in being able to answer faulty science questions by detecting the incorrectness in the question.

Method: This is a concept of **Multi Agent Systems**. 2 LLMs are set up. These are 2 agents essentially. Each row in the dataset is fed to the Agent 1, which is responsible for answering the faulty science question. The response is then passed to the LLM as a judge.

The LLM as a judge which is another agent has been set with system prompts to detect flaws in the question or evaluate the response of the 1st agent. This agent will then provide feedback to the 1st agent on how to proceed with answering the question with hints. **This is called ReAct Architecture = Reasoning + Action.**

This method on an average produces better results with LLMs to answer questions that can detect faults in science questions

Question: ***Can human-in-the-loop systems improve LLMs' detection and correction of flaws in real-time?***

Answer: *Yes, with LLM as a judge / human in the loop, better responses to discipline specific faulty questions can be obtained and the LLM will be better equipped to handle faulty science questions. LLMs can also be improved using **Chain of Thought Prompting**.*

References

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