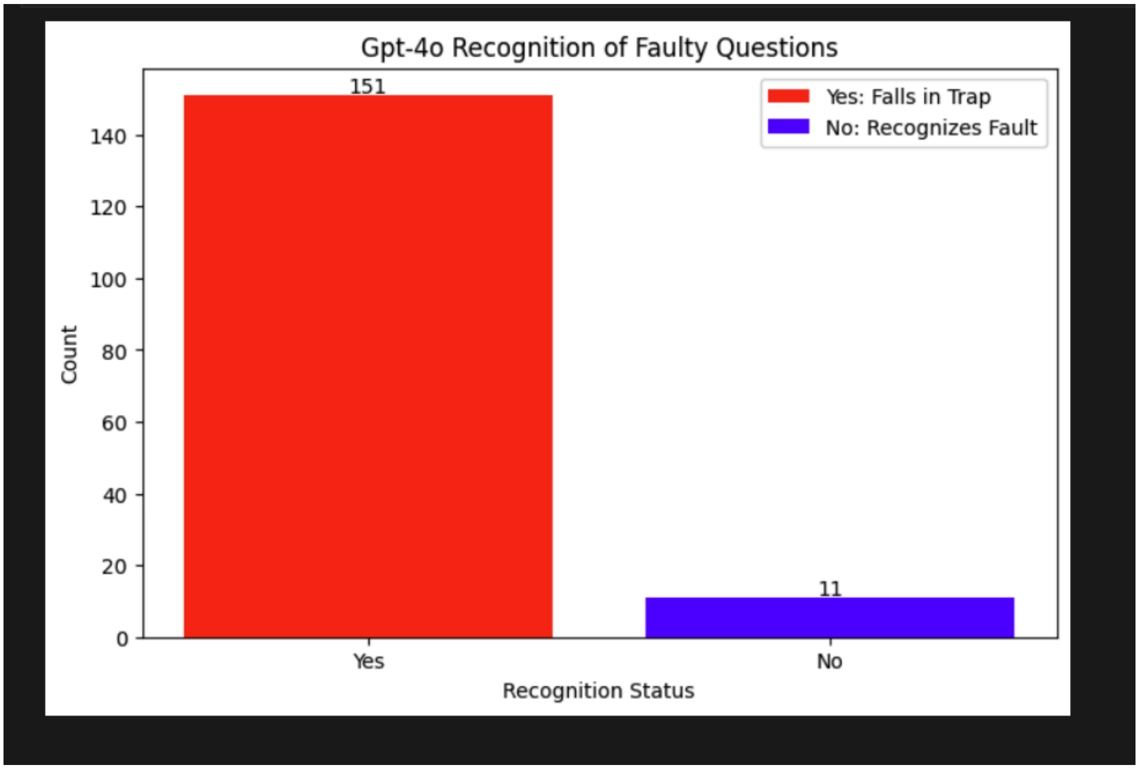
Faulty Science Questions — Research Questions (Final Project)

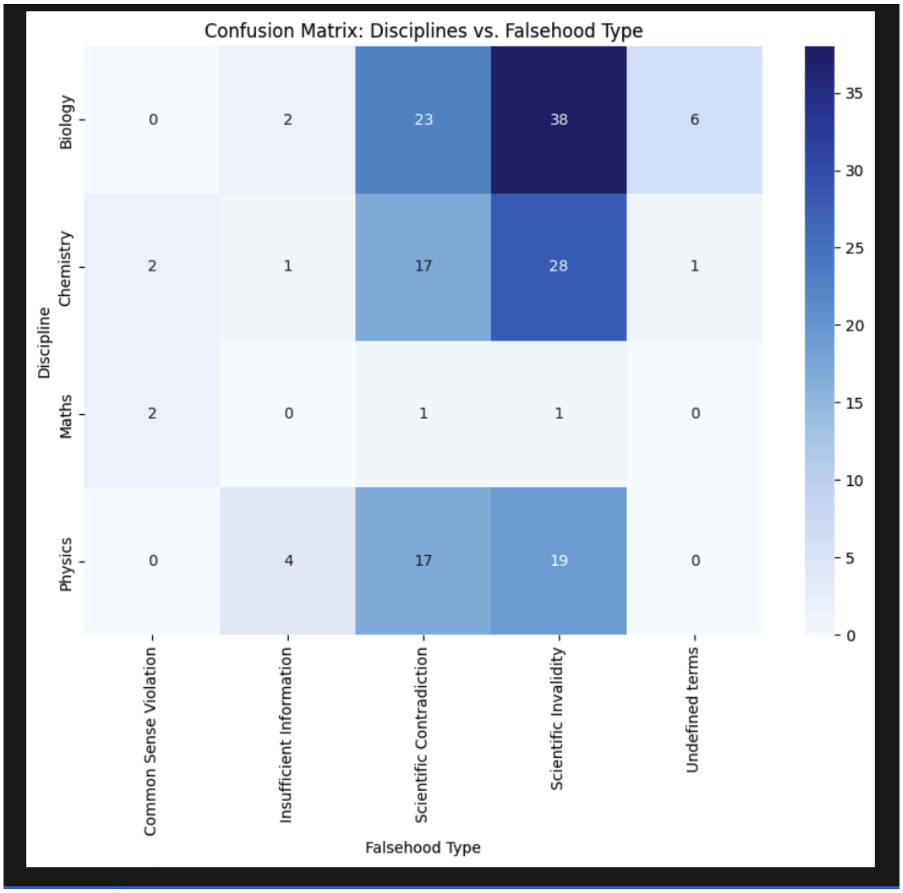
Q1. How effectively can an LLM recognize and evaluate the reasoning flaws in responses generated by another LLM when addressing faulty scientific questions?

Finding an answer to this research question can help us understand how well LLMs can perform as fault detectors for other AI-generated outputs. If LLMs are to serve as a self-evaluative tool or peer-reviewers in systems, they should possess the ability to detect and reason through faults. This can help us ensure that AI output, especially in the domain such as science and education, are reliable and accurate. With this motivation, I performed some experiments. Let's assume that LLM_1 is the LLM used to generate an answer for the faulty question and LLM_2 acts as an evaluator which will evaluate the answer of LLM_1. Here, LLM_2 was prompted and informed that the questions are faulty and the response from LLM_1 may or may not realize this. Upon experimenting on a dataset of 162 questions, 151 (93.2%) questions were flagged by LLM_2 as “falling into trap” while only 11 (6.8%) questions were successfully identified and accounted for the fault. This experiment indicates that even when the LLMs are given explicit prompts to assess the reasoning flaw, LLMs can struggle to validate logical flaws. For this experiment, both LLM_1 and LLM_2 were GPT-4o. Future experiments can check if different fault types such as logical inconsistencies, scientific invalidity, or factual inaccuracies are easier for LLMs to detect than others. This can help us understand the areas where LLMs struggle to reason. Another experiment can be to modify the prompt and ask it to explain the reason for choosing “yes’ or"no ". Studies have shown that introducing such prompts where LLMs are asked to reason before coming to a conclusion, have shown improved results. The below bar graph visualizes the results discussed above.



Q2.Which types of faults in scientific questions are most challenging for LLMs to identify, and how does this vary across disciplines with uneven representation?

This research question aims to find the types of fault that LLMs struggle with the most. For this experiment, I categorized the 162 faulty science question dataset into seven different fault types viz. Common sense violation, Insufficient Information, Scientific Contradiction, Scientific Invalidity, Misleading Statement, No Solution, and Undefined Terms. Since, the dataset had 4 disciplines viz. Biology, Chemistry, Maths and Physics, the distribution of falsehood was studied across all 4. The dataset distribution across different disciplines was uneven (eg: Biology had more questions than MAths). Despite having an uneven distribution, the results reveal that for faulty questions pertaining to Scientific Invalidity and Scientific Contradiction, the LLMs fail to identify the problem. This suggests that the reasoning flaws for such domains are difficult for the LLM to identify, indicating the fundamental limitation in the LLMs ability to reason across such faults. On the other hand, when we have more explicit faults like common sense violation or undefined terms, they are easily detected. These findings can be worked on further to see if there are any underlying reasons behind why LLMs fail to detect such violations. Below is a confusion matrix that shows the representation of falsehood across all the 4 disciplines that I have in my dataset.



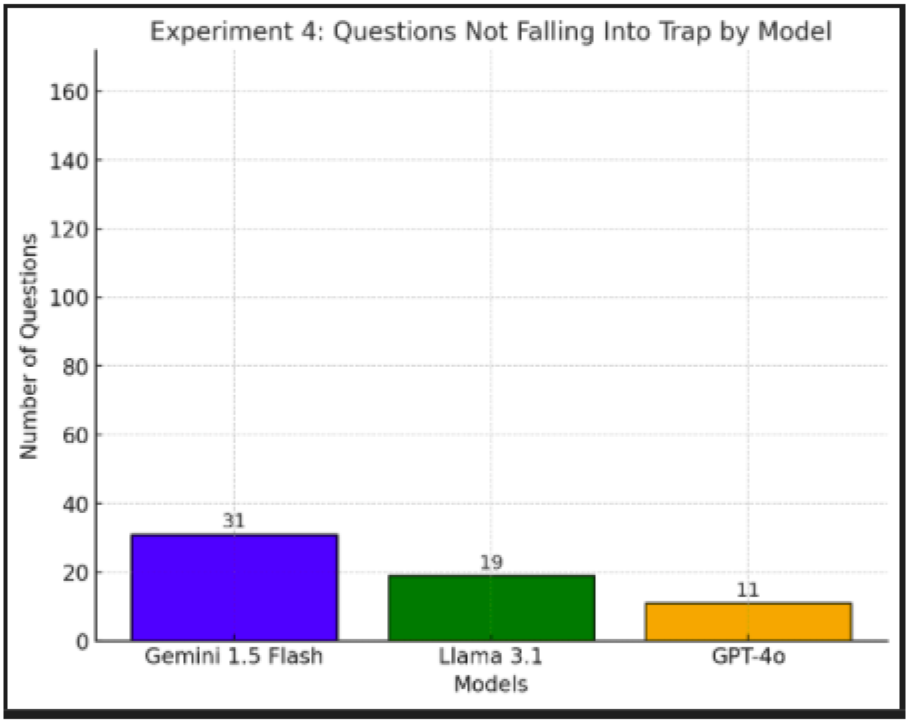
Q3. How does the framing of hints (true, false, or no hint) influence an LLM's ability to detect faults in scientific questions?

The research question helps explore whether providing contextual hints can affect the LLMs ability to identify faulty questions. To find answer to this question, an experiment was conducted where the LLM was asked to reason under three different conditions — (1) no hint was provided to solve the question, (2) a true hint was provided stating that there is a possibility of the question being factually incorrect and hence unsolvable and (3) a false hint was provided stating that all questions are valid solvable questions. The experiment was conducted on the same dataset of 162 faulty questions, and the results were compared to the baseline when no hints were given. In the baseline, the LLM was able to identify 11 (6.8%) questions as faulty. However, when a true hint was provided, the number of identified faulty questions increased to 52, which is nearly 32% of the dataset. This shows that there was an increase of nearly 25% when the LLM was prompted with a hint. On the other hand, when a false hint was provided, the number of identified faulty questions increased to 47, which is nearly 29%. This increase was slightly lesser than the increase when a true hint was provided. This experiment indicates that hinting whether true or false, improves the LLMs reasoning ability. The true hint had a higher success rate than the false hint, suggesting that LLMs are influenced by how the prompts are framed and how explicit cues can activate critical reasoning processes. The results of this experiment aligns with the famous chain of thought prompting technique. The small difference between true and false hints (52 v/s 47) suggest that LLM does not strongly differentiate between correct and misleading context, indicating some input bias. The results of this experiment reveal that LLMs can benefit from prompts that encourage skepticism or critical evaluation, which is especially important in fields like science, where factual accuracy is important. The below table provides accuracy of Gpt-4o when passed with no hints, true hints, false hints. The accuracy indicates the number of questions it was able to identify as faulty.

Method	**LLM falls in Trap**	**LLM identifies the fault**
Gpt 4o + No Hints	151 (93.2%)	11 (6.8%)
Gpt 4o + True Hints	110 (67.9%)	52 (32.1%)
Gpt 4o + False Hints	115 (70.9%)	47 (29.1%)

Q4. How do different LLMs compare in their ability to detect and reason through faulty scientific questions?

To evaluate the fault-detection capability of different LLMs, experiments were conducted on three LLMs — Gemini 1.5 Flash, Llama 3.1 and Gpt 4o. These experiments were on the same dataset of 162 faulty questions. Each model was given the faulty questions to solve and the results were human analyzed. Across all three models, Gemini 1.5 flash demonstrated superior performance. The responses from Gemini were more robust in handling faulty problems. Out of 162 questions, it falls in trap for 131 questions which is better than our baseline of gpt 4-o (where 152 questions were answered). Llama 3.1 fell into the trap for 143 questions (88.3%), indicating moderate performance but a noticeable gap compared to Gemini. Gpt-4o was the weakest, falling into the trap for 151 out of 162 questions (93.2%). These results highlight the clear disparities in fault-detection capabilities among LLMs. Gemini 1.5 Flash's superior performance suggests that it is better at reasoning or validating questions, potentially stemming from the training data or fine-tuning process used to develop it. In contrast, Gpt-4o had the worst performance, seems more prone to over-trusting the input question's framing and solving it without evaluating the validity of the problem. This experiment shows us that not all LLMs are equally capable of handling flawed input and the architecture along with data on which the model was trained significantly impacts its ability for such use cases. The below graph helps us visualize the results from this experiment.


Experiment 4: Questions Not Falling Into Trap by Model
Number of Questions
Gemini 1.5 Flash
Llama 3.1
GPT-4o
Models