CSE584 Final Project

Krishnasai Paleti - kxp5619

December 2024

1 Dataset Creation

To build the dataset, I gathered multiple-choice questions from diverse sources across six disciplines: mathematics, physics, chemistry, biology, geology, and astronomy with logical reasoning(6 x 50 = 300 questions). These questions were modified to create faulty scenarios by intentionally removing the correct answer from the provided options and replacing it with an incorrect one. This transformation ensured that each question inherently lacked a valid answer among the choices, making it a faulty question.

The modified questions were then passed to the gpt4o-mini model without any additional prompts. The model's responses were recorded and analyzed. The key observation was that the LLM consistently selected one of the provided incorrect options or the one closest to the correct answer, instead of identifying and stating that none of the given options was correct. This behavior demonstrates a significant flaw in the reasoning capabilities of the LLM, as it failed to recognize the absence of a valid answer.

All questions for which the LLM provided a faulty response were included in the dataset. The primary reasons for these faulty responses can be attributed to:

- Hallucination: The LLM fabricated or assumed an incorrect answer.
- Improper Reasoning: The LLM failed to logically evaluate the absence of a correct answer among the provided options.

Once the dataset was compiled, I formulated several research questions, focusing on how LLM performance varies across disciplines and how specific prompt designs impact its reasoning abilities. These research questions guided the design of experiments aimed at analyzing the susceptibility of the LLM to faulty reasoning in each domain and exploring techniques to mitigate this issue.

2 Research Questions and Experiments

2.1 RQ1: Whether explicitly instructing the LLM to identify if the correct answer is missing improves response accuracy.

Design an explicit prompt: Add explicit instructions, e.g., "*If none of the options are correct, respond with 'The correct answer is missing.*"

- Send all questions to gpt4o-mini with explicit prompt and record responses for analysis.
- Measure how often the explicit prompt improves identification of missing answers. Add a **Result** column to the dataset and based on the response, assign 1 if LLM was able to correctly identify that the question is faulty, 0 otherwise and plot the graph to visualize.

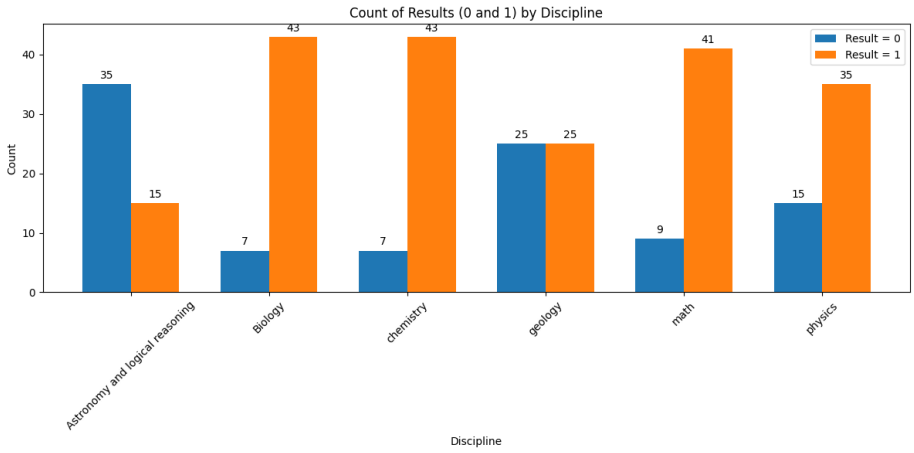

Figure 1: Analysis of RQ1

Based on fig 1, we can infer that except geology and astronomy with logical reasoning, every other discipline had a good improvement from the default pass.

2.2 RQ2: Whether encouraging step-by-step reasoning improves the LLM's ability to identify missing correct answers.

Design Step-by-Step Prompts: Include detailed instructions encouraging logical breakdown., e.g., *You are a helpful assistant. Answer the following question. Please reason step by step and identify if the correct answer is missing. If the correct answer is missing, say 'MISSING"*

- Use the same dataset and send all questions with the step-by-step prompt.
- Compare the LLM's reasoning with the step-by-step prompt to responses with default or explicit prompts. Add a **Result** column to the dataset and based on the response, assign 1 if LLM was able to correctly identify that the question is faulty, 0 otherwise and plot the graph to visualize.

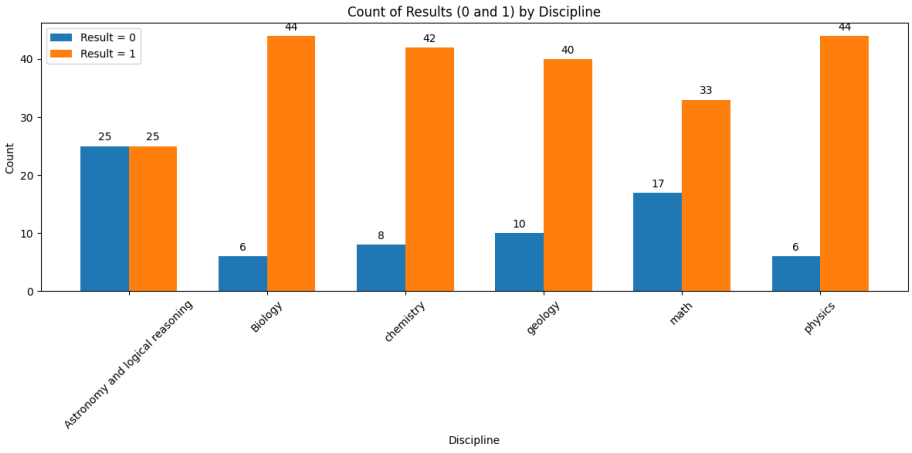

Figure 2: Analysis of RQ2

Based on fig 2, we can infer that except astronomy with logical reasoning, every other discipline had a good improvement from the default pass. Math is another that suffers in comparision when we do a chain of thought reasoning(step by step reasoning). We can infer that logical reasoning when mixed with a companion discipline such as astronomy, LLM has a hard time to reason step by step.

2.3 RQ3: Whether few-shot examples improve the LLM's ability to identify missing correct answers.

Design Few-Shot Examples:Create 2–3 sample questions with missing answers and include the correct LLM response., e.g., *Q1: What is the atomic number of helium? Options: 1, 3, 5. Correct Answer: The correct answer is missing. Q2: What is the boiling point of water at 1 atm? Options: 99°C, 100.5°C, 101°C. Correct Answer: The correct answer is missing. Now, answer the following question, if there no correct answer, say 'MISSING'*

- Use few-shot prompts with all questions in the dataset
- Compare few-shot responses to zero-shot (default) responses. Add a **Result** column to the dataset and based on the response, assign 1 if LLM

was able to correctly identify that the question is faulty, 0 otherwise and plot the graph to visualize.

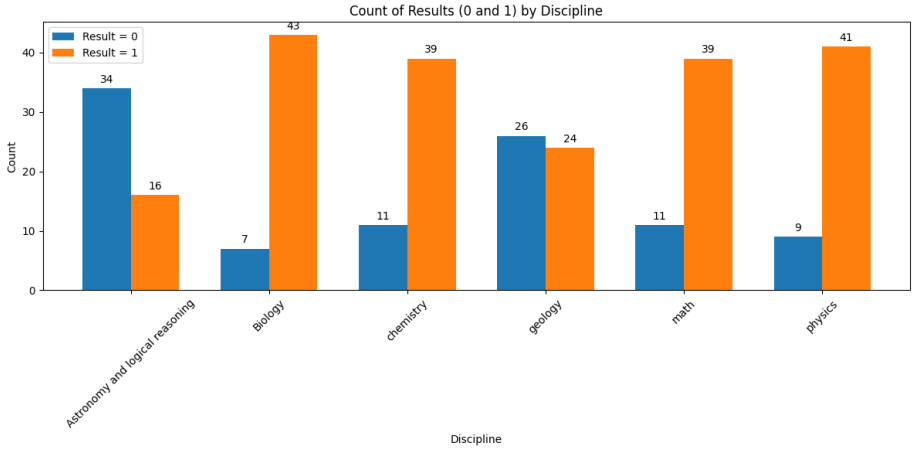

Figure 3: Analysis of RQ3

Based on fig 3, we can infer that except geology and astronomy with logical reasoning, every other discipline had a good improvement from the default pass. This shows that transfer learning via few shot examples is not as successful as core science domains such as math, physics, chemistry and biology. Thus a likely conclusion would be to experiment more and be more targeted in few show examples specifically for geology and astronomy.

2.4 RQ4: Whether reframing the question as a binary task improves LLM accuracy (e.g., "Is the correct answer missing?")

Create Reframed Prompts:e.g., *Is the correct answer to the following question missing? only answer 'yes' or 'no'*

- Use the binary prompts for all questions.
- Measure whether the LLM performs better on a binary classification task (missing or not missing) than selecting an option. Add a **Result** column to the dataset and based on the response, assign 1 if LLM was able to correctly identify that the question is faulty, 0 otherwise and plot the graph to visualize.

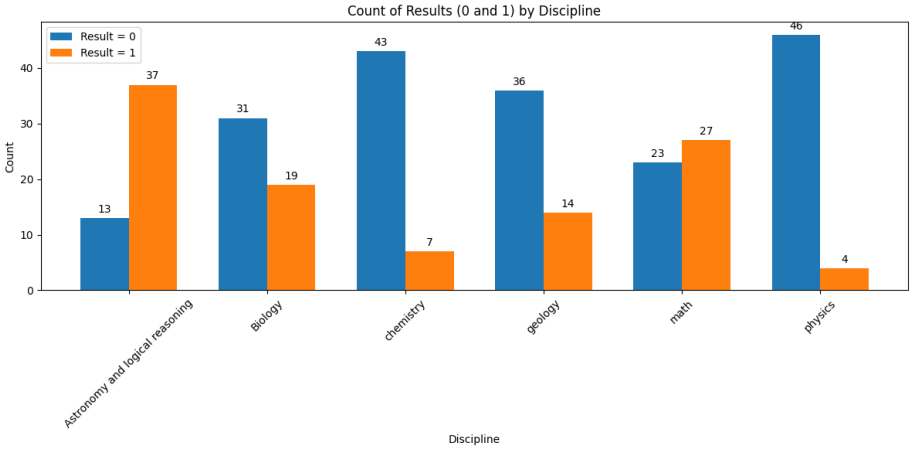

Figure 4: Analysis of RQ4

Based on fig 4, we can infer that astronomy with logical reasoning, every other discipline had a worse performance to default pass. This shows that logical reasoning questions perform well when there is a binary prompt. Every other discipline suffered and the LLM was not able to identify if the answer existed or not.

2.5 RQ5: Whether asking the LLM to compare and contrast options helps it recognize missing answers

Create Contrastive Prompts:e.g., *Consider the following question from different perspectives. Compare each option carefully with the correct answer to decide if the correct answer is missing. If the answer is missing, say 'MISSING'*

- Apply these contrastive prompts to the dataset.
- Assess whether the additional comparison step leads to better accuracy in recognizing missing answers. Add a **Result** column to the dataset and based on the response, assign 1 if LLM was able to correctly identify that the question is faulty, 0 otherwise and plot the graph to visualize.

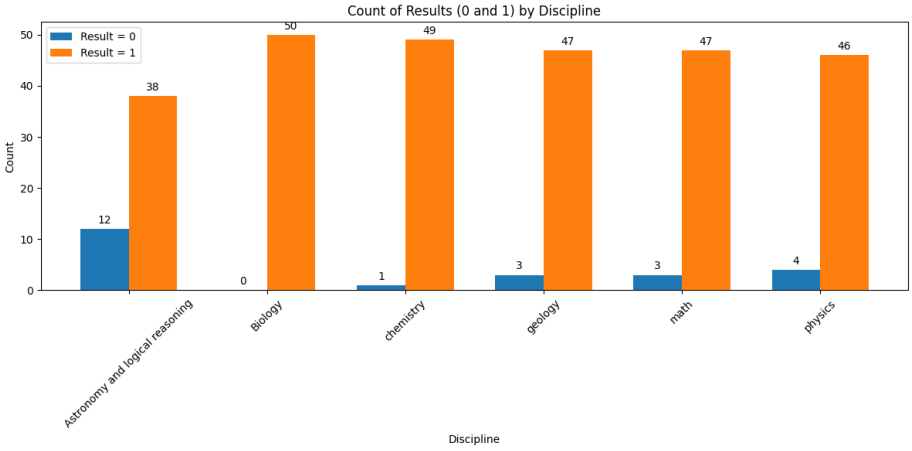

Figure 5: Analysis for RQ5

Based on fig 5, we can infer that this prompting strategy and this type of prompt is the best among all 5 tried. Most of the disciplines had perfect(biology) or near perfect detection by the LLM. In essence we need to prompt the LLM to add an additional comparison step to avoid choosing from the available or closest options.

3 Conclusion

Research Question 5(RQ5) has been the best prompting strategy in the experiments conducted. RQ5 also had the highest success rate for the LLM to say there is no answer available in the MCQ type faulty questions.