CSE-584 Final Report

Introduction & RQs:

As Large Language Models (LLMs) become increasingly popular among the general public and researchers alike, we are interested in exploring how effectively the current top-performing LLMs handle reasoning tasks and resolve scientific questions. To investigate this, we constructed a small dataset containing faulty science questions designed to mislead LLMs (We found some of them online). These questions exploit logical inconsistencies, aiming to test whether the models directly compute answers without recognizing the underlying issues within the questions themselves.

We identified couple interesting research questions and designed experiments to address them, including:

- *RQ1*: How effectively can current top-performing LLMs detect and explain logical inconsistencies in faulty science questions?
- *RQ2*: Can providing additional context or rephrasing faulty questions improve the LLM's ability to identify and correct logical inconsistencies?

Additionally, we came up a few interesting research questions that we didn't have sufficient time to explore:

- **Human vs. LLM Reasoning**: Since LLMs aim to emulate human-like reasoning as part of achieving artificial intelligence, it is critical to identify the exact differences between pre-trained models and human reasoning capabilities.
- **Comparing Performance LLMs on Reasoning**: By collecting results from different LLMs, we can determine which model perform best in addressing reasoning-based questions.
- **Impact of Training Data:** Exploring how training data influences the reasoning capabilities of LLMs can provide valuable insights. Understanding how these models solve reasoning tasks and the extent to which training data affects their performance is an important avenue for future research.

Experiments:

RQ1: To evaluate how well LLMs can identify and explain logical inconsistencies in faulty science questions, we compute the detection rate of the logical typo across all faulty questions and group by questions types as table below:

Discipline	Baseline Detection	Count	Identifies Issue count
Astronomy	Missed issue	1	0
Biology	Missed issue	4	0
Chemisty	Missed issue	1	0
Counting	Missed issue	2	0
Ecology	Missed issue	1	0
Genetics	Missed issue	1	0
Geology	Missed issue	1	0
History	Missed issue	1	0
Linguistic	Missed issue	3	0
Math	Missed issue	3	0
Physics	Missed issue	29	1
Relational	Missed issue	2	0
Science	Missed issue	1	0
Spatial	Missed issue	7	0
Thermodynamics	Missed issue	1	0

However, due to the progress of the top-performing LLMs, at the time we submit this report, most of top-performing LLMs are able to detect logical issues or typo from the faulty questions.

RQ2: Since all top-performing LLMs were able to detect the logical inconsistencies or typos in the faulty science questions at the time of conducting this experiment, we were unable to evaluate their limitations in handling such scenarios.

Conclusion:

Through our experiments, we thought to address two primary research questions: the effectiveness of LLMs in detecting logical flaws (RQ1) and the potential for improved performance with rephrased or contextualized questions (RQ2). While the models demonstrated strong detection capabilities for logical inconsistencies or typos in the dataset, this limited our ability to evaluate LLMs' weaknesses or identify scenarios where they might fail.

Our findings underscore the robustness of top-performing LLMs but also highlights the need for further investigation into their reasoning mechanism, particularly in comparison to human cognitive processes. Future work should delve deeper into comparing reasoning capabilities across LLMs, understanding the influence of training data on performance, and exploring how these models can better emulate nuanced human reasoning.