Problem Statement

Large Language Models (LLMs) such as Chat GPT, Claude, and Gemini are becoming increasingly used for question answering. While LLMs have demonstrated fairly acceptable accuracy with simple questions, the use of LLMs for answering questions in higher academic and research settings can be risky. LLMs are known to sometimes produce factually incorrect or misleading information with high confidence, raising significant concerns when used as educational or scholarly tools.

This project is inspired by the TruthfulQA benchmarking project and aims to assess LLMs' ability to answer higher level math questions. The models will be probed without providing explicit examples. Trends in model errors and limitations will be analyzed.

Data Sources

The primary dataset will be generated by prompting multiple LLMs with a set of complex competition math questions sourced from the Omni-MATH dataset found here: https://omni-math.github.io/. These questions are competition-level problems covering topics like algebra, geometry, and number theory, and each problem includes step-by-step solutions. Each dataset entry will include the model's answer and a label for correctness.

Methods, Techniques, and Technologies

1. Dataset Preparation
    - Creating subset of questions that spans across various difficulty levels and categories
2. Model Evaluation
    - Query LLMs (ChatGPT, Claude, and Gemini) via APIs and/or manual prompting based on cost
    - Use zero-shot prompting and temperature = 0 for consistency
3. Labeling Responses
    - Check mathematical equivalence
    - Run semantic similarity between responses and dataset's solutions
4. Analysis
    - Compute accuracy
    - Qualitative analysis of common error types (e.g., confident but wrong, calculation errors, misunderstanding problems)

Deliverables

1. **Code:** notebook with code
2. **Dataset:** Subset of ~350 math questions in CSV format, including topic and difficulty annotations.
3. **Evaluation Script:** Code for prompting the models, recording outputs, and scoring truthfulness.

4. **Labeled Response Dataset:** Model answers with labels for correctness.
5. **Analysis Report:**
    - Metrics (accuracy)
    - Tables summarizing model performance
    - Example correct and incorrect responses
    - Observations and conclusions about LLM math truthfulness