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 aims to investigate the relationship between confidence and correctness in LLM-generated responses to advanced academic questions, specifically in STEM fields. We will label the responses with the following labels: 1) confident and correct, 2) confident and incorrect, 3) not confident and correct, or 4) not confident and incorrect. Then, we will train a model to predict these outcome categories based on the features of the responses – such as phrasing, tone, and linguistic markers – to evaluate whether we can detect confidently wrong information.

Data Sources

The primary dataset will be generated by prompting multiple LLMs with a set of complex academic questions sourced from the MATH dataset found here: https://github.com/hendrycks/math. 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, the model's self-rated confidence level, and a label for correctness.

Methods, Techniques, and Technologies

Several tools and techniques will be used to accomplish this project's goals. First off, we need LLMs to generate the data. Chat GPT will be prompted with questions manually due to API cost restrictions, while Gemini will be prompted through its free API. Prompt engineering will be used to elicit answers and confidence levels. Based on the nature of the answers (e.g. numerical versus natural text), we may use text analysis to discern the semantic similarity between the correct answer and the given answer. Next, we will use a classification model to use features like word choice, sentiment, length, certainty phrases, and model metadata to classify outputs into the four categories. The model will be evaluated with accuracy, F1-score, and confusion matrices. Correlation between reported confidence and actual correctness will be assessed.

Deliverables

1. **Dataset**: a structured of LLM responses to academic questions with labels for correctness and confidence level
2. **Analysis Notebook:** data analysis on response patterns and trends between confidence and correctness
3. **Classification Model:** trained model that can predict whether a given response is confidently correct, confidently wrong, etc.

4. **Evaluation Report:** summarizing findings, evaluation metrics, limitations, and suggestions for future research
5. **Presentation:** video presentation on project goals, methodology, results, and insights