Final Project - Progress Update 1

Diya Vinod

April 15, 2025

04/15/25 Updates
Completed Tasks:

1. Changed data collection methodology in proposal
2. Wrote code to get controversial questions from the DELPHI dataset
3. Decided on LLM prompts

Tasks I still have left:

1. Input prompts into various LLM Models
2. Calculate the semantic entropy
3. Compare the semantic entropy of controversial prompts vs noncontroversial prompts

I have yet to conduct any experimental results as I have been focusing on data collection for the past week. However, I plan to begin experimenting this week.

The code used for generating the dataset can be found in the github or in the proposal document.

I changed how I was planning on classifying the prompts based on the feedback I received from my peers on my proposal. In the feedback I received, they mentioned that the way I was categorizing the prompts into fact-based vs biased questions was not defined enough. Therefore, as recommended by one of peers, I looked into the DELPHI dataset. This dataset expands on the Quora Question Pairs Dataset. I used these questions to determine the controversial and noncontroversial questions to ask the LLM. By using this dataset, I have used the methods used in the paper to classify my prompts.

In addition, I have few more ideas of factors that I would like to explore. It would be interesting to see how temperature affects the semantic entropy. Also looking at reasoning models to see how their semantic entropy differs from the common LLM could be interesting as well. If I have time I might expand upon my calculations and see if I can determine some type of threshold that can be used to determine if a model is hallucinating or not. For example, a user could calculate the semantic entropy of their output and if it is higher than the threshold they would need to exercise more caution because it highly possible the model is hallucinating.