CS162 Course Project: Midterm Report

Introduction

As large language models continue to grow and be used in extremely influential spheres, concerns have risen about their ability to spread inaccuracies and unfair biases in communication. It becomes increasingly important to optimize a model's ability to differentiate factual claims from unfactual, and unfair claims from fair claims. If a model can be trusted to evaluate claims correctly, it can be a tool to mitigate human uncertainty in real life scenarios.

This project is designed to provide us an in-depth understanding of the generative paradigm in Natural Language Processing. Our main objective is to determine "to what extent can LLMs accurately detect the factuality and fairness of language?" Through hands-on experience with models like Phi-2 and the UnilC benchmark, we can explore the extent to which LLMs can recognize and ensure the factuality and fairness of their outputs.

We are given milestones which range from evaluating the models ability to classify claims based on fairness and factuality, processing data for model prompting, and employing advanced prompting techniques to enhance classification. These milestones allow us to explore the essential components needed to build an NLG.

Methods

We used the Phi2 LLM model and used the train_claims and test_claims provided to us to test the model straight out of the box. At first, we were experimenting with an untrained PhiForCausalLM model which resulted in our output consisting of nonsensical terms. After experimenting with AutoModelForCausalLM with the Phi2 version, we found much better

results on the training and testing data. Our best submission test link would be under Trial 5 in Week 7.

Results & Discussion

Accuracy on Test Set: 0.7

F1 Score on Test Set: 0.67

To address the current limitations and further improve the model's performance, we can conduct thorough error analysis on the areas where the model fails to predict specific types of inputs. In doing this, we can hope to configure the model configuration. Another way we can further improve the model performance is moving on to milestone 3 where we can experiment with different techniques to enhance the fairness and factuality classification ability of the LLM Phi-2. We can look at papers that utilize the Uni-LC dataset to find places to start research.

In addressing these issues, our goal is to surpass the baseline performance and ensure the model becomes more reliable in its predictions. We should allow time to carefully plan, experiment, and validate our tests to ensure the models overall performance and usability.