Group 19 Project Midterm Report

Noah Gardner, Joel Hernandez, Carson Cox, Minghao Chen

Introduction:

In today's world, it has become increasingly important that statements are made tactfully and without bias. As the usage of Large Language Models becomes more and more prevalent in our daily lives, their struggles with fairness and accuracy pose a quintessential issue. In an effort to discover just how much we can rely on Large Language Models to detect factually incorrect or unfair biases in language, we must study how they evaluate these factors with a variety of statements. If these models are to continue to grow in demand and usage, we must be absolutely certain they provide unbiased and fair responses.

Our project uses the Phi-2 Large Language Model in conjunction with the Unified Language Checking (UniLC) benchmark suite of statements. The UniLC benchmark contains a suite of statements concerning recent events – climate change, public health, and hate speech – on which we can test our Large Language Model's ability to discern the aforementioned factors of bias. Our main goal was to read some statements into the model we created and determine if what was being claimed was a true and fair statement, or alternatively, an unfair or biased assumption.

Methods:

We didn't end up experimenting too much with datasets or models beyond the basics of completing milestone 2 with zero_shot evaluation. We attempted to run it with Flash Attention 2 as instructed, but had some compatibility issues. We had to instantiate the model without it, and thankfully it still worked perfectly fine. We implemented the dataset class capable of transforming the UniLC benchmark statements into prompts suitable for the Phi-2 model, utilizing a zero-shot evaluation prompt template. We utilized the provided PHI_ZERO_SHOT_EVAL from constants.py when testing our milestone 2 code on the test set. The zero_shot eval tested the Phi-2 model's ability to discern bias and inaccuracies against the varied topics in the test file. The configurations, including batch size and tokenizer settings, were chosen to optimize model performance and evaluation accuracy to pass the threshold of 67.2% acc and 67.4% f1. We also mostly implemented the other prompt types in dataset.py, but didn't get around to fully testing them beyond the dummy claim yet.

Results & Discussion:

Our Phi-2 model was able to achieve a performance of 0.72 accuracy and 0.67 f1_score on the test set, which we submitted to Gradescope under "Project Test Set Trial-4 (Week 7)". This surpasses the baseline given to us, which highlights the potential of

LLMs to navigate through complex bias and factual verification in long pieces of text. But it also highlights the opportunities for further optimization in model performance and evaluation techniques that we will look into in the future for milestone 3 and further open-ended experimentation. Through batch prompt processing, we were able to evaluate the Phi-2 model's scalability and efficiency. This method highlighted the model's ability to maintain consistency across a wide range of prompts while using parallel processing to gain better throughput. Single prompt focused more on understanding and looking at the response accuracy of individual queries, showcasing its adeptness at capturing nuances and context with good precision.

References:

https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/