Final Executive Summary: Super GLUE LLM Benchmark Classification

Amber Walker, Andrew Bennett, Caroline Belka, Luis Alvarez

January 21, 2024

We selected the SuperGLUE LLM benchmark dataset for our project. Specifically, we chose the BoolQ dataset which consists of a yes or no question, and a passage that contains information about the question. The label is the answer to the question either yes or no.

We began by doing exploratory analysis on the data and created a random classifier which gained around the expected 50%. We then continued by creating a rules based classifier. We identified a few patterns within the dataset that included identifying negation terms such as "this is NOT the case" which would be an indicator of a no answer. The distribution of the negation word counts was high and sparse and our results were worse than 50%, so we decided on a new avenue of question and passage completeness. The idea was that the more words from the question were in the passage, the more correct the answer would be. This technique gave us higher accuracy, but it was still poor. In the final iteration, we decided to subset the data into a specific amount of questions that compared two things, like "is this the same as that." We were able to properly classify these statements by using subset rules logic to achieve greater than 50% accuracy.

In the process of creating the rules logic, we determined that it was incredibly difficult to get context from False Positives and False Negatives. We then created the BERT classifier with 32 labels, which performed better than our rules based classifiers. We continued with zero-shot, data augmentation techniques, and LLM data. For our data augmentation technique, we used combinatorial math to create new subsets of the same data, but there was extra noise and we got scores of slightly better than 50%. Because of the noise, we paraphrased the data and created a new subset which improved performance. Our LLM data was generated by both GPT3 and GPT4, and the model trained with GPT4 performed better than the GPT3 trained model. Finally, because we did not have much improvement from any of our techniques, we decided to modify the core of the training data with masks. Additionally, because of the sequential nature of the data we implemented an RNN and saw close to 70% accuracy.

Then we began training BERT on incremental data loads. We saw increased accuracy results from 1% to 50%, but surprisingly saw a plateau of performance from 50-100%. This may be due to the repetitive nature of the data, and the subtle nuance of the data that we have continually mentioned. We then compare the training techniques while training with a full dataset. The results interestingly showed the combined LLM training and the masked RNN training being the most accurate model, and slightly outperform the subset trained models. We again could not find significant patterns in the False Positives or False Negatives because of the similarity and nuance of the data.

We then distilled the model and successfully increased speed from 282 to 242 seconds but at a slight cost to accuracy, which dipped by 4% to 62.25%. This trade-off between speed and precision is the key of distillation, where the temperature and alpha hyperparameters play crucial roles. Higher temperature values can enhance the student model's nuanced understanding of classes, while the alpha parameter adjusts the weight between teacher-influenced soft labels and actual data labels. For improvements we consider an attention-based distillation for a more refined knowledge transfer, which would also require parameter tuning to balance efficiency with accuracy.

In summary, we were able to gain insight on the SuperGLUE benchmark, understanding some of the issues and nuances of the dataset. We implemented multiple new training techniques, and used LLMs to augment our data. Overall, the BoolQ dataset is difficult to understand and given more time and resources, a more in depth analysis of question level nuanced may prove to increase accuracy. However, we had a in depth exploratory analysis and used many differing techniques to try to gain insight into the nuances of the dataset.