Role of Student vs. Mentor

I conducted all the work for this project since it was entirely CS and didn't require a lab or additional assistance. Dr. Kevin Crowthers was a mentor for me during this project, guiding progress and helping me establish general directions.

Equipment and Materials

Python 3 and Visual Studio Code were the programming language and development environments used for this project. HuggingFace, a popular machine learning platform that houses various open datasets and pre-trained transformer architectures, was used to download datasets and models. PyTorch, a machine learning framework, was used in conjunction with HuggingFace for all training, inference, and arithmetic operations. Other minor software packages used include Python libraries, such as Matplotlib, Seaborn, and other data processing libraries. The hardware requirements include the WPI Compute Cluster, whose immense storage and computing power were necessary for large-scale data operations and model training.

Dataset Analysis

WikiText, a commonly used LLM dataset, was downloaded as raw compressed data from HuggingFace. Using various Python scripts and techniques for subset analysis, the data was analyzed with respect to various risk assessment variables, including sentence completion, lexical diversity, Flesch–Kincaid score, etc. Using Seaborn and Matplotlib, distributions of cross-dataset scores were visualized. After quantification and testing across the dataset, the targeted risk-assessment variable was narrowed and isolated for mitigation with a data cleaning technique.

Data Cleaning

Again, using Python and packages such as Pandas and other NLP libraries, the targeted risk-assessment variable had a corresponding data cleaning technique developed to attenuate it. Using the WPI Compute Cluster, the algorithm was applied across the WikiText datasets, creating a cleaned version that used heuristic filtering to sieve training examples classified as low-quality. Once the dataset was cleaned, it was used for training.

Model Training

GPT-2, an LLM model architecture, was loaded from the HuggingFace Transformers library and used for training with the raw and clean datasets. Using Python, PyTorch, and the HuggingFace documentation, the model was then trained on both datasets with the aid of the WPI Compute Cluster. After training, the models trained on the different datasets were benchmarked using BIG-bench and various task-specific LLM metrics. The improvement in model performance after data cleaning was then characterized using statistical tests.

Statistical Tests

A z-test was used in this project using mean projections derived from BIG-bench to establish the statistical significance of relative model improvement due to data cleaning. A z-test was used since the metric output can be characterized as approximately normally distributed.

Z-test

This section would be a section underneath/included under the Statistical Tests Secondary Header. This sub-section is labeled with a tertiary level (Level 3) format. You should not include the equation for the statistical test unless it was developed as a key component of your project. Equations should be formatted accordingly:

$$a^2 + b^2 = c^2 \tag{1}$$

The equation should be labeled numerically surrounded by parentheses, and towards the right margin. Equations should also be referenced in the text (Equation 1).