Course: CSE 584 Modified
Date: October 6th, 2024
Participant: Xinzhang Xiong

Data Curation

This project develops a classifier to distinguish between text completions generated by three different Large Language Models (LLMs): GPT-4, Llama 3, and Gemini 1.5 Pro. The classifier, trained on a crafted dataset of 2,370 data points equally from the three LLMs, given the same truncated pieces of Xi.

In the data file '20240928project1_data2', first column contains the label where:

- 1 is from GPT-4
- 2 is from Llama 3（Meta AI）
- 3 is from Gemini 1.5 Pro

And the second column contains the completed sentences by the corresponding LLMs. With this dataset, the model has 2862 tokens and 283203 parameters.

The prompt "I want you to finish each sentence with at least 20 words. Please make them complete sentences and unique with each other. Please keep the order as given." was fed into LLMs for obtaining the data.

Classifier and Architecture

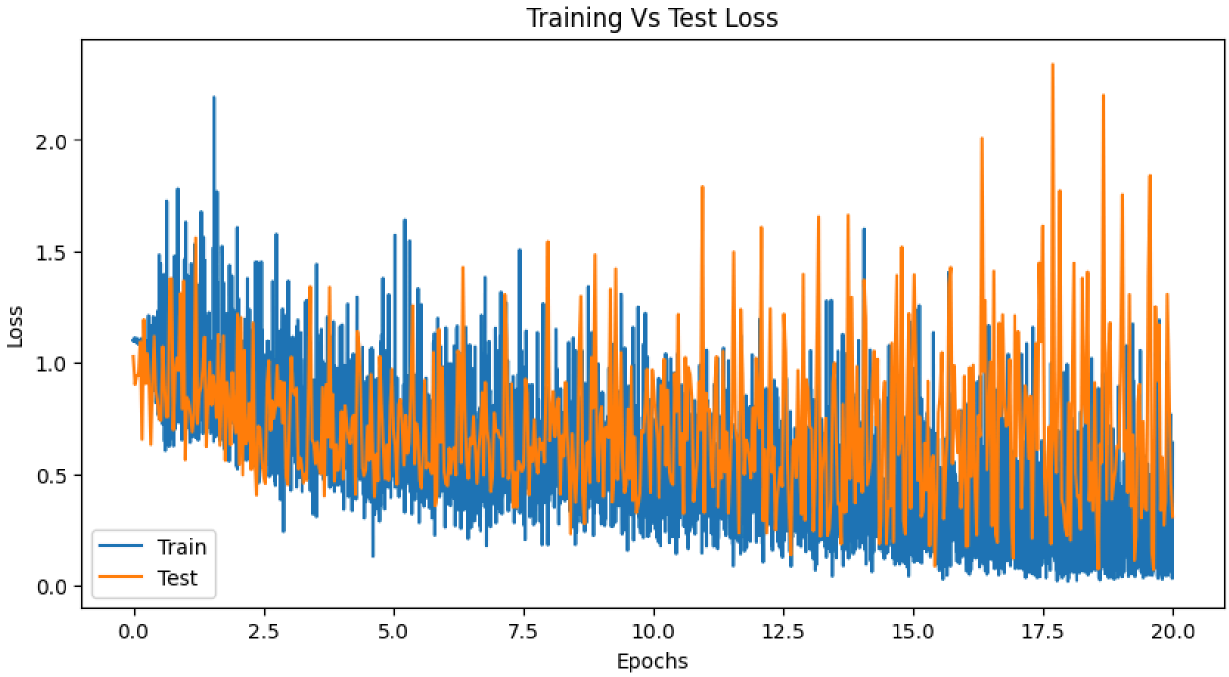
These parameters were fine-tuned during experimentation to achieve optimal performance training using the developed small dataset size:

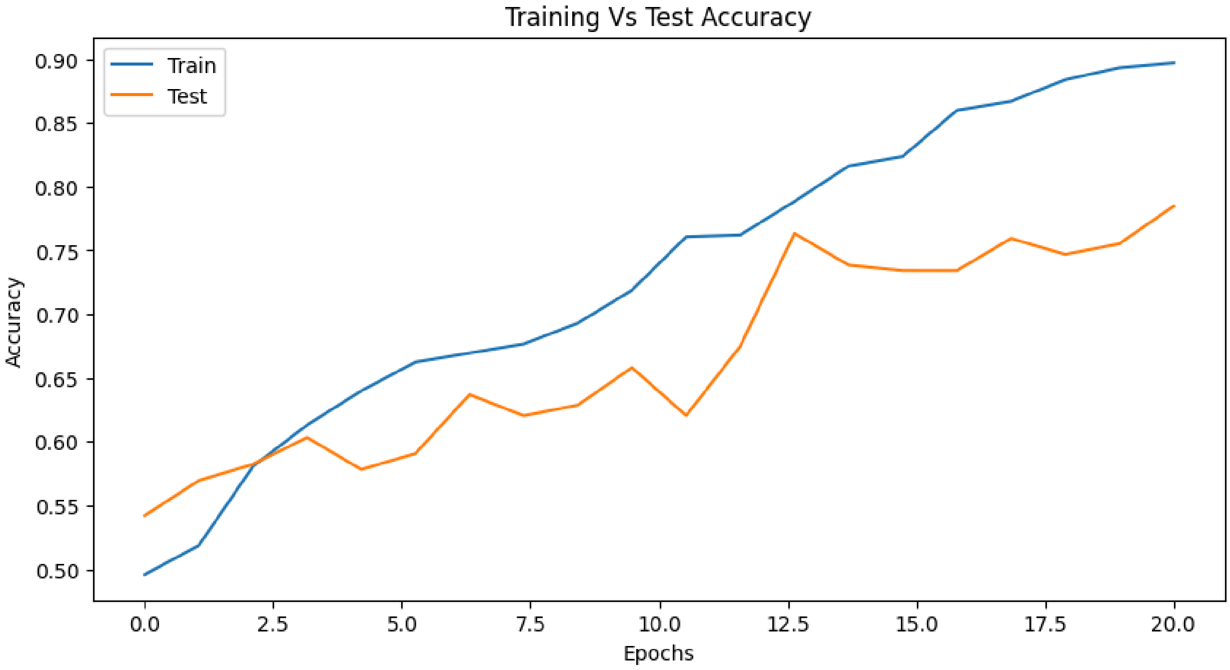
- **Dropout:** Set to 0.1 to prevent overfitting while maintaining a balance with the small data size.
- **LSTM Layers:** Only one LSTM layer was used, with a hidden size of 128 to control model complexity.
- **Learning Rate:** 2e-4
- **Batch Size:** 8
- **# epochs:** 20
- **Maximum length for a input sentence:** 64

The LSTM model processed sequences of tokens, including start (<sos>), end (<eos>) markers, handling unknown (<unk>) words and padding (<pad>) to normalize sentence lengths, and learned to predict the label corresponding to the LLM that generated the completion.

Results

Despite the small dataset, the model achieved a best test accuracy of over 78%. This performance was reached after careful tuning of hyperparameters. However, the model tended to overfit the training data, given its limited data size.





As depicted in the top figure, the training loss steadily decreases, indicating ongoing learning. However, the loss for the test dataset stabilizes at a certain level, suggesting a convergence point. Similarly, the bottom figure shows an increase in training accuracy, while the test dataset accuracy begins to plateau. Extensive experiments with up to 50 epochs demonstrate that this trend of overfitting worsens as training continues.

In-depth Analysis

This project highlighted the potential and limitations of using LSTM classifiers for different LLM output classification tasks. Strategies for choosing appropriate parameters for the given data set were used by combining theoretical understanding and testing for a different range of parameters, including dropout, number of LSTM layers, learning rate, batch size, number of epochs, etc.

On the other hand, when retrieving the data from LLMs, it was observed that LLMs struggle with maintaining the order and accuracy of sentence completions when prompted with long or complex text inputs. Modifying prompt instructions over time revealed that increasing the required completion length allowed for richer data for classification but introduced more variability in LLM outputs**.**

Related Work

Recent advancements in LLMs have primarily focused on improving the quality of text generation, with limited attention to the classification of model outputs. Prior studies have utilized various features of text generation for model identification but have not extensively explored the use of LSTM classifiers for this purpose. This project contributes to this niche but growing area of research.

Text classification: A perspective of deep learning methods

Z Wan

arXiv preprint arXiv:2309.13761, 2023·arxiv.org

```
@article{wan2023text,
  title={Text classification: A perspective of deep learning methods},
  author={Wan, Zhongwei},
  journal={arXiv preprint arXiv:2309.13761},
  year={2023}
}
```

Deep learning--based text classification: a comprehensive review

S Minaee, N Kalchbrenner, E Cambria, N Nikzad, M Chenaghlu, J Gao

ACM computing surveys (CSUR), 2021·dl.acm.org

```
@article{minaee2021deep,
  title={Deep learning--based text classification: a comprehensive review},
  author={Minaee, Shervin and Kalchbrenner, Nal and Cambria, Erik and Nikzad, Narjes
and Chenaghlu, Meysam and Gao, Jianfeng},
  journal={ACM computing surveys (CSUR)},
  volume={54},
  number={3},
  pages={1--40},
  year={2021},
  publisher={ACM New York, NY, USA}
}
```