COE 379 Final Project Proposal

Goal

With the lowering costs of machine learning models there are more LLM models, available now more than ever. For this technology to become widely adopted across different industries, it needs to be fine tuned for each specific use case.

The goal of this project is to fine tune an existing LLM to speak like different people. We plan focusing on artists since there is a large availability of data on the internet. The artist we emulate in the project will be the rapper "Chief Keef". Specifically we are focusing on thematic continuation where we give the ml model a verse and continue the next line like chief keef. While we are focusing on artists in this project, this concept can be applied to numerous different fields such as healthcare, finance, entertainment, etc.

Data Sources to be used

In order to train the model, we are planning to use song lyrics available from kaggle as well as web scraping from az lyrics for any missing artists. Then we will use this data to fine tune a LLM model with 67M parameters. We believe that this model is trained enough to "chat" with but has enough flexibility to be fine tuned

Artist lyrics

- https://www.kaggle.com/datasets/deepshah16/song-lyrics-dataset?resource=download
- https://www.azlyrics.com/c/chiefkeef.html

LLM to fine tune
- https://huggingface.co/distilbert/distilbert-base-uncased-finetuned-sst-2-english

Methods/Techniques we are considering

We are planning on fine tuning the existing llm with a training dataset we provide.

Specifically, we are planning on splitting our dataset by different lines within songs. For example, the training model will take an even take sets of lines (2,4,6…) and split into input and output sets. This will simulate a user feeding a line and the model continuing it. We can test the accuracy of the verse generation using a **BLEU scoring method**. While traditionally used for translation methods, this can be used to compare phrase syntax similarity.

The model would be given the test data inputs and predict the next line. Then the generated lines would be compared to the true lyric using the BLUE scoring method.

Product Deliverables

The final deliverable would be a model uploaded to hugging face that allows users to input a verse of their choice, and generate the next line as if chief keef was singing.

If the model creation goes well, some other parameters we would like to test would be general questions. The model will not be fine tuned to these parameters, but rather used to understand the biases.

There will be 3 categories of questions we plan on asking…

- Direct Style Imitation Questions
  - Testing the style of language the model uses.
  - Example question "What are some things that Chief Keef might say about trust and loyalty in his crew?"
- Reaction and Response Questions
  - These questions simulate conversations that the artist might have.
  - "What would Chief Keef say if he were asked about his biggest influences in music during an interview?"
- Creative and Abstract Questions
  - These questions stretch the models ability to reason to unknown contexts while maintaining their style/
  - "How would chief keef write a song about aliens in space?"