Capstone Proposal - Fine-tuning Llama3 with Harry Potter dataset

Domain Background

Machine learning has revolutionized natural language processing (NLP), enabling the creation of intelligent chatbots that understand and generate human-like text. Open-source models like Llama 3 have made advanced NLP capabilities accessible to a wider audience. Combining these capabilities with a popular and engaging dataset like the Harry Potter series provides an exciting opportunity to create educational tools that make learning English fun and interactive.

Harry Potter, written by J.K. Rowling, is one of the most beloved series of all time, with rich language and intricate plots that captivate readers. Leveraging this series for educational purposes can motivate learners by immersing them in a familiar and enjoyable context.

Problem Statement

Learning English, particularly as a second language, can be challenging and often lacks engaging and interactive methods. Traditional learning approaches may not capture the interest of learners, making it difficult for them to stay motivated and practice regularly. There's a need for an engaging, interactive tool that can provide instant feedback and make the learning process enjoyable.

Solution Statement

I propose to fine-tune the Llama 3 open-source model using a dataset derived from the Harry Potter series to create a chatbot. This chatbot will ask users questions about Harry Potter, engaging them in conversations related to the movies. Additionally, it will correct users' mistakes, providing explanations to enhance their learning experience. This interactive approach will not only make learning English more enjoyable but also provide practical language usage practice.

Datasets and Inputs

The project will utilize the **Harry Potter Movies Dataset**, which contains extracted text from all seven Harry Potter movies, focusing on dialogues and descriptive passages to train the model in understanding and generating relevant content.

Benchmark Model

As a benchmark, we will start with the pre-trained Llama 3 model. This model, already proficient in general language understanding, will be fine-tuned with our specific datasets to improve its performance in the context of Harry Potter-themed interactions and error correction.

Evaluation Metrics

To evaluate the performance of the fine-tuned chatbot, we will use the following metrics:

1. **Perplexity**: Measures how well the model predicts a sample.
2. **BLEU Score**: Evaluates the quality of text generated by the model by comparing it to reference texts.

Presentation

The final deliverable will be a user-friendly chatbot interface accessible via an Android application. Users will be able to interact with the chatbot, answer questions about Harry Potter, and receive feedback on their English language usage. The project report and a demonstration video will be provided to showcase the model's capabilities and the chatbot's performance.

Project Design

The project will be executed in the following stages:

1. **Data Collection and Preprocessing**:
    - Extract text data from the Harry Potter movies.
    - Collect and preprocess the English language learning dataset.
    - Tokenize and clean the text data for model training.
2. **Model Fine-Tuning**:
    - Fine-tune the Llama 3 model with the Harry Potter text corpus.
3. **Chatbot Development**:
    - Develop the chatbot interface using Android technologies.
4. **Evaluation and Testing**:
    - Evaluate the model using the defined metrics.
5. **Deployment and Presentation**:
    - Deploy the chatbot on an AWS Endpoint
    - Prepare the project report and demonstration video.

By following this structured approach, we aim to create an effective and engaging educational tool that leverages the popularity of Harry Potter to enhance the English learning experience.