Project Proposal:

Title: Developing a BERT Model for Technical Literacy Classification

Team Members: Tagore Kosireddy, Mihret Kemal, Feven Tefera

Introduction:

Our project aims to train the BERT model using the technical literacy dataset derived from the Yahoo questions and answers dataset. With the consent of Dr. Evan Lucas, we'll incorporate research conducted by one of our team members, Tagore, into this project.

Dataset Collection and Preprocessing:

The project involves collecting, cleaning, and preprocessing a large dataset, specifically targeting technical literacy-related content. We'll employ various prompt engineering techniques to assign appropriate scores to the data on a given scale.

Prompt Engineering and Model Selection:

Additionally, we'll explore different large language models to determine which one performs best with these techniques. Once we've created the technical literacy dataset, we'll convert the scores into labels based on predefined ranges and encode them accordingly.

Model Training and Evaluation:

Using the BERT model, we'll train it using techniques such as cross-validation, ensuring the dataset's balance and employing the Adam optimizer with cross-entropy loss. To monitor the model's performance, we'll plot loss curves for training, validation, and testing, aiming to avoid overfitting or underfitting. We'll also assess the model's predictive accuracy using various performance metrics.

Exploration of Alternative Models:

Time permitting, we'll explore smaller BERT models to see if they offer comparable performance. Additionally, we'll conduct a literature review on BERT-related research to gain insights into its workings and best practices.

Project Goals and Deliverables:

Ultimately, our goal is to develop a BERT model capable of accurately classifying technical literacy data. We plan to create an interface for this model, making it accessible for users to interact with and explore its capabilities.