Consensus Technical Assessment: Language Model Fine-Tuning, Evaluation, and Optimization

Overview:

This assessment is designed to evaluate your skills in fine-tuning a language model and optimizing it for efficient, high-quality performance. You are provided a small training dataset (training_data.csv), and your goal is to provide a model that performs the following task, with some latency parameters. A second inference dataset (inference_data.csv) is intended to measure the speed at which you can compute the generated answer.

Language model task: the dataset you are provided contains questions a user may ask Consensus, and a (potentially) related paper intended to answer the question, and some additional metadata. Your goal is to build a model that takes in that question + paper, and outputs an answer to the question based only on the context in the paper.

Requirements:

1. **Speed Requirement**:
    - After fine-tuning, the model should be able to perform inference on a batch of 20 text inputs in **under 3 seconds**. This means that your approach should focus on making the model lightweight and efficient without compromising too much on performance quality. Use whatever computing resources you have available, and that can dictate the specifics of the architecture you choose.
2. **Evaluation Metrics**:
    - This model is intended to answer questions from scientific papers, and you are provided with a label of the expected answer. You should provide a metric, or series of metrics, that give a holistic picture of how well the model performs its task

Deliverables:

1. **Private GitHub Repository,** which should contain (share with GitHub user **bnebeker**):
    a. **Training Code**: any relevant code to data preparation, training, saving of the model
    b. **Inference Code**: Code that demonstrates inference on a batch of 20 inputs, ensuring it meets the 3-second requirement.
2. **Metrics Definition + Justification + Outputs:** what metrics you chose, why those metrics are appropriate for this task, and the actual calculated values
    a. Because of the latency requirements, the “why” is more important than the metric itself