**Course Project Progress Report**

**1. Team Members:**

- Kun Peng: Responsible for overall coordination, data analysis, and model development.

- Lining Yu: In charge of model fine-tuning and evaluation.

**2. Progress:**

**Dataset description:** [MedQuAD](https://github.com/abachaa/MedQuAD) is a dataset designed for medical question answering tasks. It comprises a collection of questions sourced from various medical domains, accompanied by relevant context passages and corresponding answers. This dataset includes 47,457 medical question-answer pairs created from 12 NIH websites (e.g. cancer.gov, niddk.nih.gov, GARD, MedlinePlus Health Topics). The collection covers 37 question types (e.g. Treatment, Diagnosis, Side Effects) associated with diseases, drugs and other medical entities such as tests.

**Basic models:**

The BERT (Bidirectional Encoder Representations from Transformers) dataset in NLP (Natural Language Processing) refers to a corpus of text data used for pretraining BERT models.

**Preliminary results:**

Data preprocessing

Training model with subset of dataset



**3. Problem:**

Limited resource for computation.

**4. Backup Plan:**

For backup plans, if BERT does not yield satisfactory results, we will explore alternative models like RoBERTa or GPT-3 for their potential in capturing the nuances of medical texts..

**5. Timeline:**

- Weeks 5-6: Model fine-tuning and initial evaluation.

- Weeks 7-8: Model adjustments based on feedback and re-evaluation.

- Week 9: Final evaluations and preparations for the Status Report.