Group Project Proposal

Team Members:

- Shikha Kumari
- Kalyani Vinayagam Sivasundari

Question - Generating System

What problem did you select and why did you select it?

We selected the problem of question answering using the SQuAD (Stanford Question Answering Dataset) dataset. We chose this problem because question generating is a fundamental task in natural language understanding, and SQuAD is a widely used benchmark dataset for this task. It consists of questions posed on a set of Wikipedia articles, with corresponding answer spans within the text.

What database/dataset will you use?

We will use the SQuAD dataset, which is freely available and widely used for question answering research and benchmarking. The dataset contains questions, passages, and corresponding answer spans from Wikipedia articles.

What NLP methods will you pick from the concept list? Will it be a classical model or will you have to customize it?

We will explore using pre-trained language models such as the T5 transformer model or similar architectures fine-tuned specifically for question answering tasks.

What packages are you planning to use? Why?

We plan to use popular NLP libraries such as Hugging Face's Transformers library for accessing pre-trained models like t5-base. PyTorch wrapper such as Pytorch lightning.

What NLP tasks will you work on?

The primary NLP task for this project will be question generating, specifically generating questions from passages given an answer.

How will you judge the performance of the model? What metrics will you use?

We will be using CrossEntropyLoss as our metric.

Provide a rough schedule for completing the project.

- **Week 1: Preprocessing & Exploration**

- **Week 2: Model Selection & Setup**
- **Week 3: Training, Fine-tuning**
- **Week 4: Populating the app in streamlit app**