COE 379L Final Project Proposal

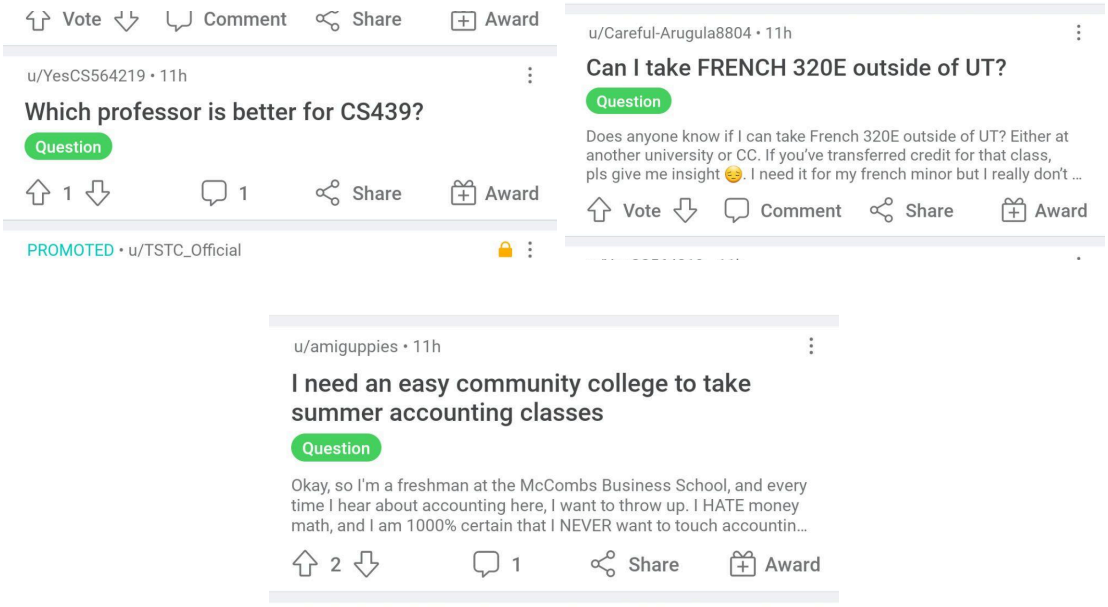

Figure 1: Screenshot of questions posted by students on the UT Austin subreddit

Current and prospective students at UT Austin often need access to a wide array of information. Moreover, some of this information can be very difficult or impossible to obtain by only using the university website or its resources. *We aim to deploy a large language model to handle question-answer queries from students using* ***Docker***, ***Flask***, ***Javascript***, ***HTML*** *and* ***CSS***. User's will then be able to communicate with this model using a client side web page written in HTML.

Data Sources

We will train our model using Reddit posts made on the UT Austin subreddit[1]. Specifically, we will train our model on posts made to ask a question–similar to those shown in Figure 1. If time permits, we may include information scraped from UT's official website, however this is unlikely given the time constraints as this would require more extensive data processing procedures.

[1]Subreddit: https://www.reddit.com/r/UTAustin/

Design and Approach

High level Design

The design of this project at a high level can be summarized using the figure below:

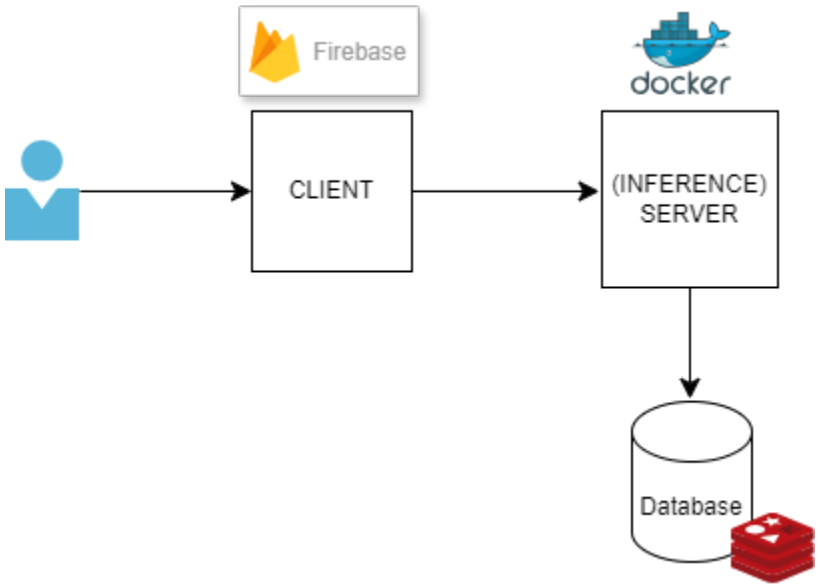

Figure 2: Project Design Diagram

Where our deployed model is run in a **Docker** container on a backend server and receives requests from the client application by users. The client application will be deployed using **Firebase** and written mostly in **React.js**. If time permits, we will also maintain a database to store user queries and improve the user experience. The database would be built using **Redis** and would be kept running in a Docker container on the backend server. , the performance of our model will be largely determined by the quality and amount of data sourced from the UT Austin subreddit.

Data Sourcing

To extract our data from the UT Austin subreddit we will explore public libraries capable of scrapping Reddit. For example, the **Universal Reddit Scrapper** is a Github repository owned by Joseph Lai, a Software Engineer at Bluestone Analytics. The Universal Reddit Scrapper (URS) includes detailed usage documentation and enables automated web scraping of subreddits (and more). Moreover, we've noticed that most question posts made by students on the /rUTAustin subreddit are tagged with a "question" label making it easier (we believe) to

filter for. Once the data is scraped from Reddit it can be saved locally in JSON format and processed before being used to train the model.

Model Design/Development

Given the limited resources for this project, the model will be built by fine-tuning a pre-trained large language machine learning model best suited for Question Answering (QA) tasks. Therefore, the pre-trained model we will use is the BERT model, as it is popular model that's often fine-tuned for domain specific QA tasks[2]. One

Deliverable

We will have two primary products: 1. A deployed inference server that handles requests from users . 2. A deployed client-side page that users can interact with to make requests and receive responses from the model. All additional work files used will be included in the Github repository of the project.

[2] Fine-tuning Strategies for Domain Specific Question Answering under Low Annotation Budget Constraints (Kunpeng et al, 2024)