COE 379L: Software Design For Responsible Intelligent Systems

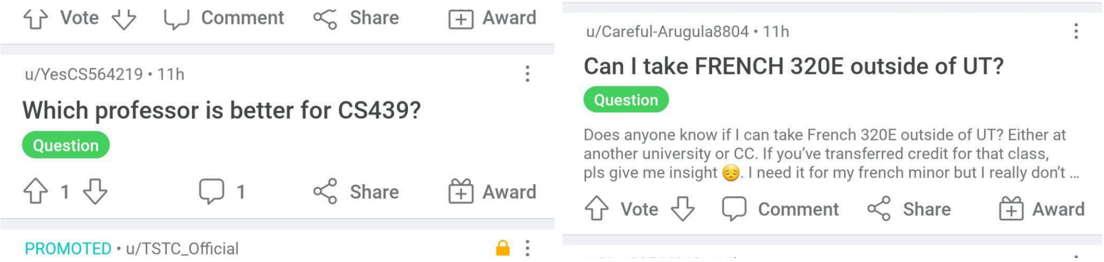
Bevo Bud The GPT

Kelechi Emeruwa and Pranjal Adhikari

May 3, 2024

Table of Contents

1 Introduction



u/YesCS564219 • 11h

Which professor is better for CS439?

u/Careful-Arugula8804 • 11h

Can I take FRENCH 320E outside of UT?

Does anyone know if I can take French 320E outside of UT? Either at another university or CC. If you've transferred credit for that class, pls give me insight 😔. I need it for my french minor but I really don't ...

PROMOTED • u/TSTC_Official

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

University students often need quick access to information such as university policies, academics, and other campus resources. However, manually searching through all university resources can be time consuming and inefficient. Furthermore, the information a student is seeking may not be available on any university sites. Various different forums and social sites, such as Reddit and GroupMe, exist where students can communicate with one another to answer these questions with anecdotal experience. Observing the large corpus of knowledge related to the university, and students' desire to access such information quickly led to the motivation for building Bevo Bud The GPT (Bevo Bud).

Bevo Bud is a full-stack web application powered by a GPT-2 model fine-tuned on a large corpus of data using the HuggingFace Transformers library. Specifically, our machine learning model was fine-tuned on data from the r/UTAustin subreddit. An explanation for why this dataset was chosen can be seen in the following section. Furthermore, in this report we aim to outline our approach in building our machine learning model while also reflecting on our result with the hopes of further improving the model. We also compare performance of our fine-tuned using two different numbers of epochs. As mentioned, we will begin with a discussion on the data used to train our model.

2 Data Source & Technologies

2.1 Data Source

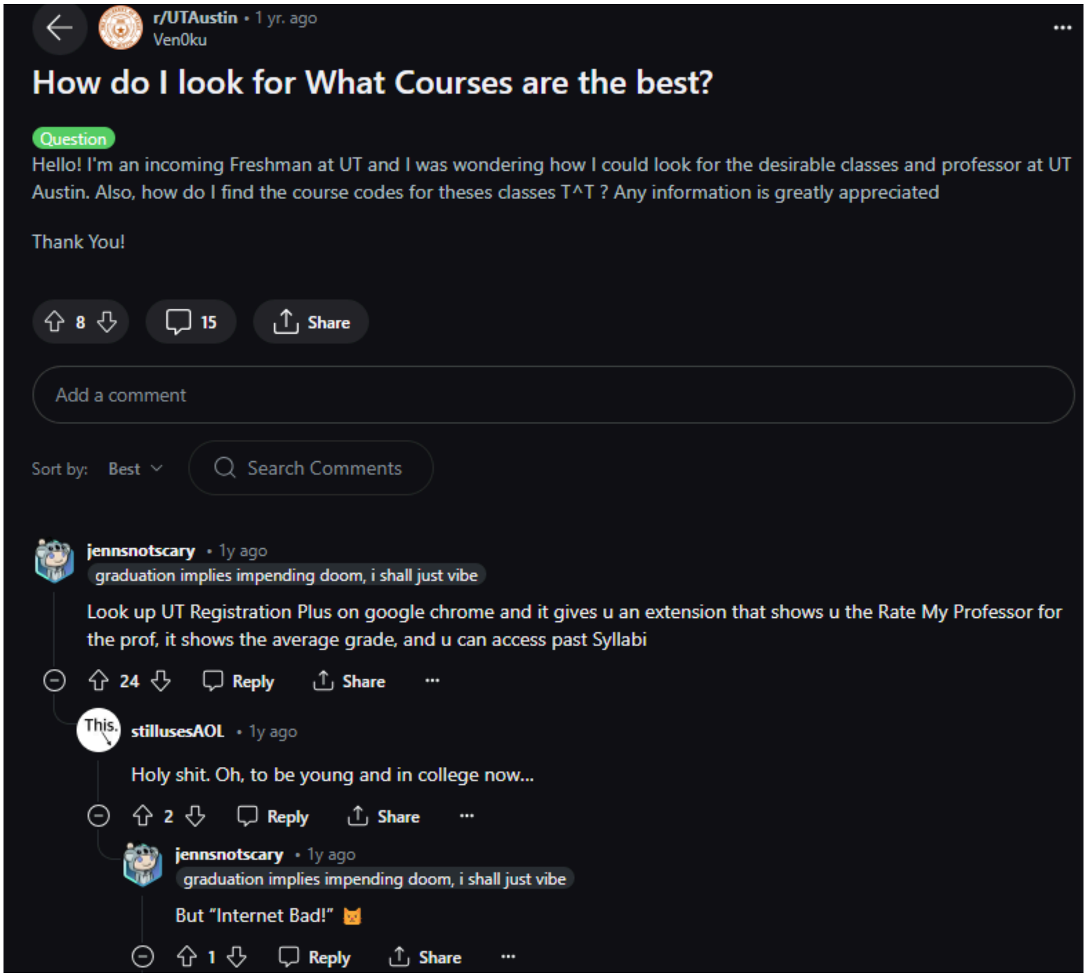
In choosing a viable data source to train our model, we were interested in finding a place on the internet where a large amount of existing data related to UT Austin already existed. Consequently, the r/UTAustin subreddit presented itself as one of our most viable options for training our model for three main reasons. Firstly, hundreds of students post questions on the subreddit *everyday*[1]. Secondly, the question-answer dialogues within this subreddit provide information that is not always accessible from an official UT Austin website[2]. Thirdly, the data was well formatted and thus easier to process and prepare. However, despite the several advantages to using the r/UTAustin subreddit to train our model, there were still issues that needed to be addressed. For example, some Reddit posts include tangential or nested conversations from users. In such scenarios, it's difficult to discern which comment qualifies as a valid answer to the initial question. The figure below is a screenshot of a Reddit post where the issue described arises. Another issue with using the r/UTAustin subreddit as training data for the model is the quality and profanity of comments from users. In fact, the same figure below demonstrates how some responses from users can be vague and/or inappropriate.

In the following section we outline how one of these limitations were addressed (although we believe the second limitation could have also been addressed with enough time). Moreover, despite the flaws inherent in relying on public user-generated data to train our model, the magnitude of the amount of raw data present in the r/UTAustin subreddit enables us to use strict filters and retain quality dialogue entries. Although using strict filters can dramatically decrease the size of the original dataset to a small fraction, a small fraction of a dataset with hundreds of thousands of entries still leaves us with (at least) several thousands of data points[3].

[1] This is based on estimates from https://subredditstats.com/r/utaustin

[2] For example, questions like "What are the best study spots on campus" are impossible to answer solely from the UT Austin website.

[3] Alternatively, we might have needed to individually generate or extract thousands of data points ourselves which would require significantly more time than programmatic filters.



r/UTAustin • 1 yr. ago
Ven0ku

How do I look for What Courses are the best?

Question

Hello! I'm an incoming Freshman at UT and I was wondering how I could look for the desirable classes and professor at UT Austin. Also, how do I find the course codes for theses classes T^T ? Any information is greatly appreciated

Thank You!

jennsnotscary • 1y ago
graduation implies impending doom, i shall just vibe

Look up UT Registration Plus on google chrome and it gives u an extension that shows u the Rate My Professor for the prof, it shows the average grade, and u can access past Syllabi

stillusesAOL • 1y ago

Holy shit. Oh, to be young and in college now...

jennsnotscary • 1y ago
graduation implies impending doom, i shall just vibe

But "Internet Bad!" 😾

Figure 2: Screenshot of a Reddit post with nested conversations and inappropriate language.

As a result, 13 million archived posts and 42 million archived comments from r/UTAustin subreddit the were downloaded[4]. After processing the data, about 11 thousand question-answer dialogue entries were used for training.

[4] Data source link: https://the-eye.eu/redarcs/.

2.1.1 Data Preparation

The data was first processed through a series of filters. The first filter served to retain only Reddit posts that were questions. Secondly, we filtered all the comments to only retain comments with a 'score'[5] greater than an arbitrary value. For our case, this arbitrary value was set to 2.[6] Next, we excluded Reddit posts that had no comments, as these could not be used to produce a question-answer entry. Lastly, to address the problem of nested conversations, we applied a filter to ignore those as well. Note, in this series of filters we do not apply filters to exclude inappropriate comments. As previously stated in Section 2.1, one way to exclude inappropriate comments could be to label such comments using classification models like HateBERT. Although this was not explored due to the time limitations, it could very easily be implemented in future iterations of this project.

Nevertheless, the resulting submission and comments were then structured into question-answer format and saved as a JSON file to help more easily fine-tune our base model. After these processing steps, there were 10,000 total unique submissions that included comments to train and develop the model.

2.2 Technologies

To design and implement the project, various technologies were used in conjunction, including Docker, Redis, Flask, Hugging Face, React, TypeScript, and Node.js. With these technologies, the model was developed and packaged for use intuitively. The design of the model can be seen in the figure below:

[5] Scores for a comment on Reddit are simply by subtracting a comments' upvote count by its downvote count.

[6] Why 2? Reddit defaults users to upvote their own comments. Therefore, filtering by a score of one is effectively the same as filtering by a score of 0. Moreover, any score above 2 may have limited or skewed our training dataset which we wanted to to avoid. With enough time we could have explored adjusting this parameter.

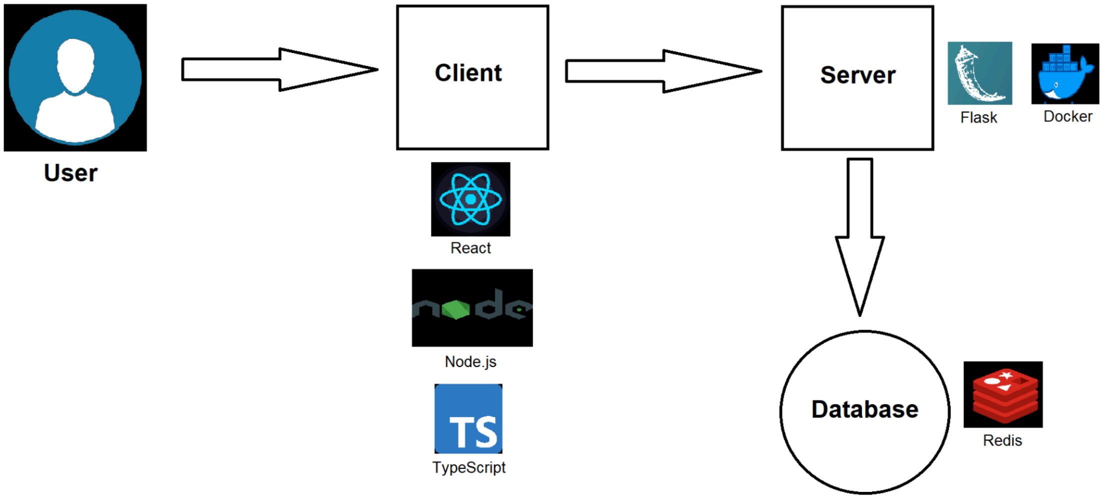


Figure 3: Bevo Bud Tech Stack Diagram

The diagram shows the service workers used in the application, each of which running on a separate docker container. The inference server loads the machine learning model as it receives requests from the user and their queries. These queries are then fed into the model for an output to the respective query. In addition, user's queries and respective output are stored in a Redis on another Docker container.[7]

React, TypeScript, and Node.js were used to develop the user side of the web application, where users will be interacting with the model. The web application is designed as an interface, allowing users to input their UT-related questions and receive a response from the model within the web interface.

Our base model was fine-tuned using the HuggingFace (HF) library. Specifically, we sought to leverage HF's Supervised-Fine-Tuning API to load and fine-tune our base GPT-2 model. Despite the existence of a large language model better suited for question and answering tasks, the GPT-2 model was chosen as our based model because of reduced overhead required to fine-tune it. For example, fine-tuning with models such as BERT would have required us to manually label thousands of data points. On the other hand, with the GPT-2 model only relatively minor re-structuring was required to

[7] The list of routes handled by the inference server can be seen in the README.md of the /server folder of the project's repository.

prepare the data. More details on how the model was fine-tuned will be discussed in the next section.

3 Methods

As mentioned, the model was built by fine-tuning a base GPT-2 model using the *SFTTrainer HuggingFace* API. Rather than use a base model suited for question and answering tasks like BERT, we were interested in fine-tuning a model well suited for causal language modeling. This is because such models are easier to fine-tune, as they do not require labeling. However, this comes at the cost of accuracy and reliability. Nevertheless, The SFTTrainer API provided high-level methods that streamline the process of fine-tuning large transformer models like GPT-2.

Moreover, due to memory constraints, we took measures to reduce the compute needed to achieve our goal. Firstly, we used a distilled version of the GPT-2 model from the Hugging Face Hub (“distilbert/distilgpt2”) as fine-tuning this model was less memory-intensive and more efficient while retaining a performance similar. Secondly, we used Low-Rank Adapters from the Hugging Face *PEFT* library to further decrease the training time needed to fine-tune our base model. Low-Rank Adapters are able to accomplish this by decomposing large matrices into smaller, low-rank matrices in the attention layers, thereby reducing the number of parameters that need to be fine-tuned, resulting in less resources and time used.

Furthermore, we built two different fine-tuned models to observe the change in performance of the model based on changes in the number of epochs. Due to length of training, we only compare two models with fine-tune training epochs of three and ten. Lastly, once our model was trained we published it onto the HuggingFace Hub where it could then be easily loaded and retrieved by our inference server.

4 Evaluation

The model fine-tuned using 3 epochs took an estimated three hours to train, and the fine-tuned model using 10 epochs took about 8 hours to train. The table below compares the performance of our fine-tuned model using 3 epochs and 10 epochs:

Table 1: Comparison of Model Responses

Question	**3 Epochs**	**10 Epochs**
How Do I Get A Parking Permit?	How Do I Get A Parking Permit? And How Would This Help? To see some of what's going on about state law, click here.	How do I get a parking permit? This is the first time that I have ever received a license. You can make the necessary purchases on a credit card with me. I don't need it because I got it,
What Are The Best Places To Study On Campus?	What are the best places to study on campus? We've got plenty of great things going on, but these are some of them: The U.K. government, one of the most important places in academia, has	What Are The Best Places To Study On Campus? The University Of Minnesota has a fantastic collection of studies looking at the state and local economies of their universities. A variety of subjects have been studied across the country, covering nearly five million Americans.
Jester East Or Jester West	Jester East or Jester West? What is 'I will save your life' and why does it matter to you if you‣n'd believe what I have said that I am doing? I don‣t	Jester East Or Jester West? Well, I guess it was a good idea to start by saying that we do not have a single "giant" community. The internet has been a little weird for some time, so some have

As we can see, the both models provide nonsensical responses to the queries related to UT Austin. There are several factors that can explain these results. Firstly, in this project we explored fine-tuning the GPT-2 model, a model that is relatively old and poor in performance. We considered using newer and more advanced models Llama 3 or GPT-Neox on the Hugging Face Hub, but memory and compute resource constraints made this infeasible. Although using the GPT-2 model as our base largely accounts for the poor performance in our fine-tuned model, there are measures also that can be taken to improve the quality of training done with our data. This includes excluding inappropriate responses, and manually revising answers.

5 Conclusion

The purpose of Bevo Bud The GPT was to help students more easily access consensus information using machine learning techniques and libraries. Although these attempts demonstrate the feasibility of fine-tuning a base model on r/UTAustin subreddit data, the poor performance of the model suggests the need for a more performant base model. Future iterations can focus on improving data quality and exploring newer models available on the Hugging Face Hub, such as Llama 3 or GPT-Neox to enhance response accuracy and relevance. Moreover, as stated previously, because the inference server loads the Bevo-Bud model from the Hugging Face Hub, the process of updating our application with newer models becomes streamlined. These steps, coupled with continued refinement of the training process, will contribute to a more effective and reliable question-answer system for students.

6 Resources

- Class website: https://coe-379l-sp24.readthedocs.io/en/latest/index.html
- UT Austin subreddit: https://www.reddit.com/r/UTAustin/
- Archived data of r/UTAustin: https://the-eye.eu/redarcs/
- HateBERT: https://huggingface.co/GroNLP/hateBERT
- HuggingFace Supervised-Fine-Tuning Trainer: https://huggingface.co/docs/trl/en/sft_trainer