Project 2

System Description

To create the machine learning model I tested out many different many different ways of achieving it. At first I tried to use the pytorch model and then the Tensorflow models shown as examples. Then I found and modeled my transformer based on the model created by Omar Khaled Ahmed on Kaggle. The model that Omar used was a encoder-decoder model with attention using LSTM architecture. I took this model and decided to use GRU as they are more efficient and less computationally intense. I tested out and adjusted some of the parameters in the processing to better suit my data and also adjusted the maximum length of the sequences to be shorter for ease and speed of training. This allowed for quicker and easier training on limited hardware. This was much needed as Kaggle only allows a limited amount of hours running with a gpu accelerator. I tried training across multiple epoch lengths ranging from 2 to 15. I also tried using different loss functions. I tried changing the metrics that the model would use to evaluate the data. I then saved both the tokenizer and the machine learning model for use with the chatbot.

For the chatbot, I reused my code from the last project. I removed the if-else statements and replaced it with a call to the model designed in the previous paragraph. This call processed the query with the aforementioned tokenizer and then ran it through the encoder-decoder. This greatly simplifies the code as it does not need to go into nested if-else statements but only one call to a model. The rest of the code is mostly unchanged. I updated the user model to be slightly better formatted but otherwise it remains the same. I also cleaned up some of the code and made it flow much better.

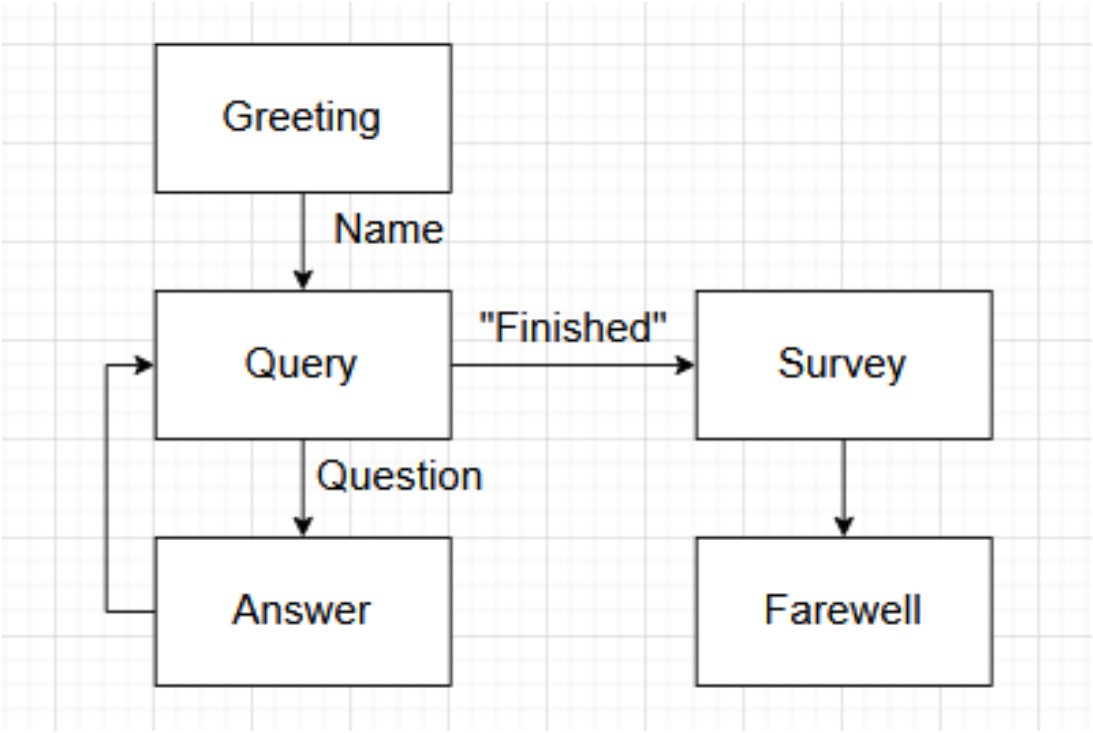
Dataset and Cleaning

The dataset that I used is from huggingface's database of datasets. I chose a dataset created by Naklecha which contained 700,000 question and answer pairs which were generated by glaive.ai using the Minecraft wiki as a source.

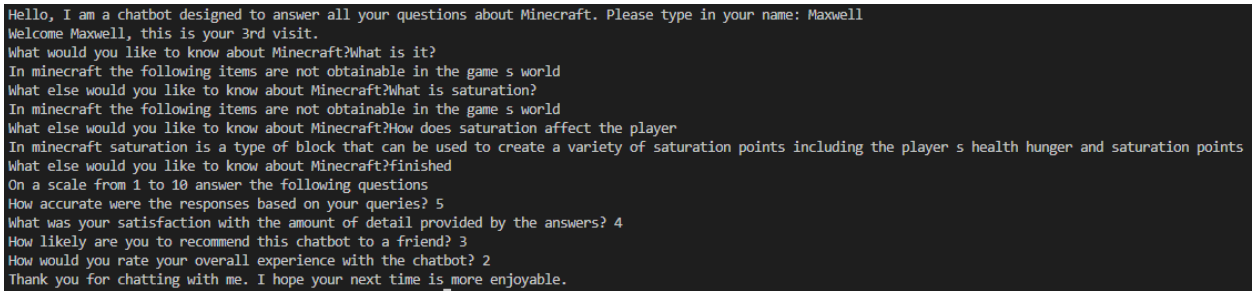
The dataset was not so well cleaned so there were a few things that I had to do to get rid of junk data. I looked through the data and found many instances where the question and answer were just the words "Question" and "Answer" or just weird in general. The first step I took was to get rid of these by filtering out questions which did not contain a question mark. This removed around 30,000 pairs from the data. I then also

filtered out any pairs where the answer was blank. This step only removed 300 entries. For the last step I created a histogram chart of the data to find out the distribution of word counts so that I could find a good place to make a cutoff for the max sequence length of the inputs for the model. The last step for the cleaning part was that I took out any pairs where one of either the question or answer was less than 2 words long or greater than 50 words long. This filtered out any questions which were just a singular question mark and answers which were not useful. It also reduced the max length of sequences so that my model would be able to train quicker. This reduced the size of the dataset by around 200,000 bringing the final size of the data to 470,000 sets of questions and answers.

Dialog tree



Sample dialog



Evaluation

This bot performs adequately at taking in questions and providing a somewhat cohesive response to the question if it is related to the questions which were provided in the dataset. It does well at creating sentences which make sense. It is not good at responding to questions that fall outside of the dataset. It will sometimes cut off abruptly if it tries to generate a sentence that is over the length of the maximum allowed word count. Overall, I think this chatbot is extremely limited in scope and usefulness.

Survey Results(1-10)

Accuracy results
4,5,6
Detail results
3,6,7
Overall results
4,6,6

Appendix

Citations

Model inspiration-
https://www.kaggle.com/code/omgits0mar/chatbot-attention-from-scratch-llms-finetune

Dataset - https://huggingface.co/datasets/naklecha/minecraft-question-answer-700k

Samples

Sample Profile

```
<Profile>
  <Name>Maxwell</Name>
<Visits>2</Visits>
<Ratings>
    <Rating1>
      <Accuracy>3</Accuracy>
    <Detail>3</Detail>
    <Recommended>3</Recommended>
    <Overall>3</Overall>
    </Rating1>
  <Rating2>
      <Accuracy>9</Accuracy>
    <Detail>9</Detail>
    <Recommended>9</Recommended>
    <Overall>9</Overall>
    </Rating2>
  </Ratings>
</Profile>
```