### 

### **Indigo Hack to Hire - Data Science (4010)**

### 

### **Report**

### 

### 

### 

### **By:**

Aditya Pandey

adpandey2102003@gmail.com

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### **Report on question-answering model**

#### **1. Introduction**

This report presents the results of an NLP project focusing on text generation tasks. The primary goal was to develop a robust NLP model capable of generating coherent and contextually accurate text based on a given input for quora question-answer dataset. The project involved data exploration, model selection and evaluation, application development, and deriving insights to recommend improvements.

#### 

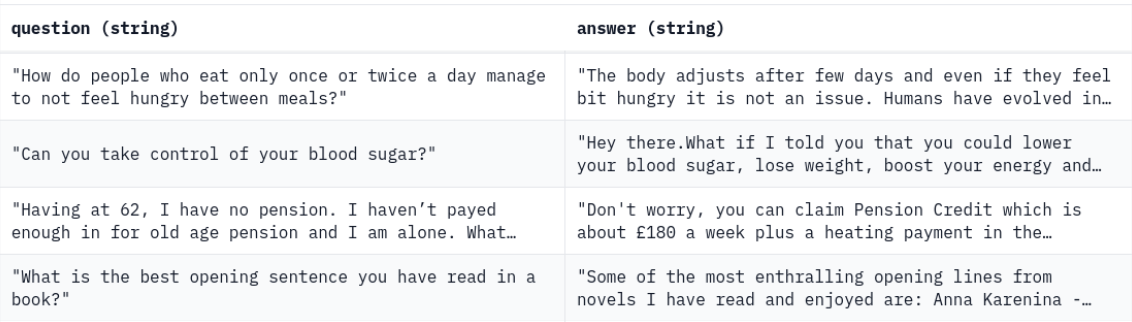
#### **2. Literature Survey**

The field of Natural Language Processing (NLP) has rapidly advanced, driven by the development of sophisticated models like BERT (Bidirectional Encoder Representations from Transformers), T5 (Text-To-Text Transfer Transformer), and GPT (Generative Pre-trained Transformer). BERT, introduced by Devlin et al. (2018), has revolutionized NLP by providing deep bidirectional representations, allowing it to understand context from both directions in text. T5, presented by Raffel et al. (2019), unified NLP tasks under a single framework by converting them into a text-to-text format, demonstrating versatility across various tasks. GPT, particularly GPT-3 by Brown et al. (2020), set new benchmarks in text generation with its capability to produce human-like language based on large-scale training data.

#### 

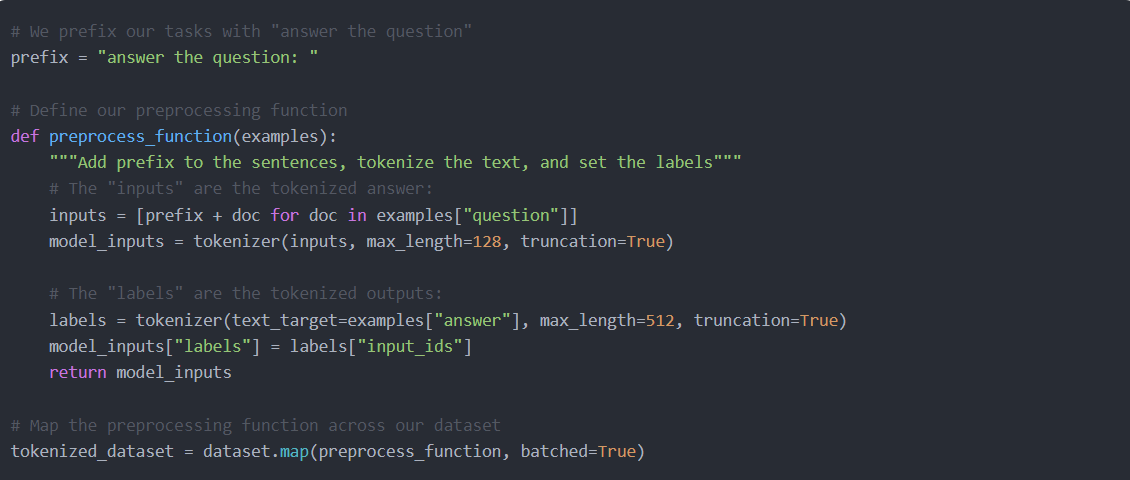
#### **3. Methodology**

**3.1 Analyze the Structure and Content of the Dataset**

The dataset used for this project consisted of textual data, which included a wide variety of sentences and passages for translation and text generation tasks. The data is in the form of question - answer.

**3.2 Data Cleaning and Preprocessing**

We use the following function to tokenize and preprocess the data. This will add the prefix "answer the question:" to all of the questions, then tokenize them. Then it will tokenize the answers. The "inputs" for training the model will be the tokenized and prefixed questions, and the "labels" will be the answers**.**

****

#### **3.3 Model Selection and Evaluation**

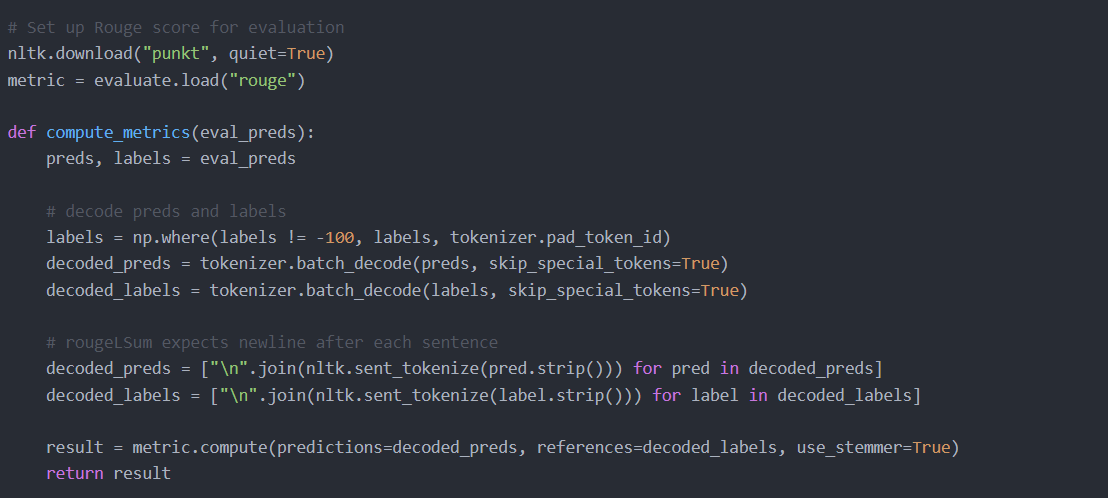
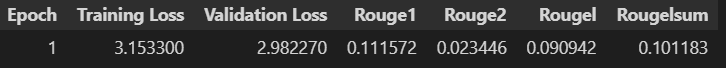
**3.4 Used NLP Models**

Two state-of-the-art NLP models were tested:

* **T5 (Text-To-Text Transfer Transformer):** A versatile model capable of various NLP tasks by converting them into a text-to-text format. This model is the one finally used in the chatbot application.
* **LLaMA 2 -7b :** A model focused on text generation with strong language understanding capabilities. Unfortunately, due to resource constraint, this model was unable to finish fine-tuning in the limited amount of GPU and ram available.

**3.5 Evaluation Metrics**

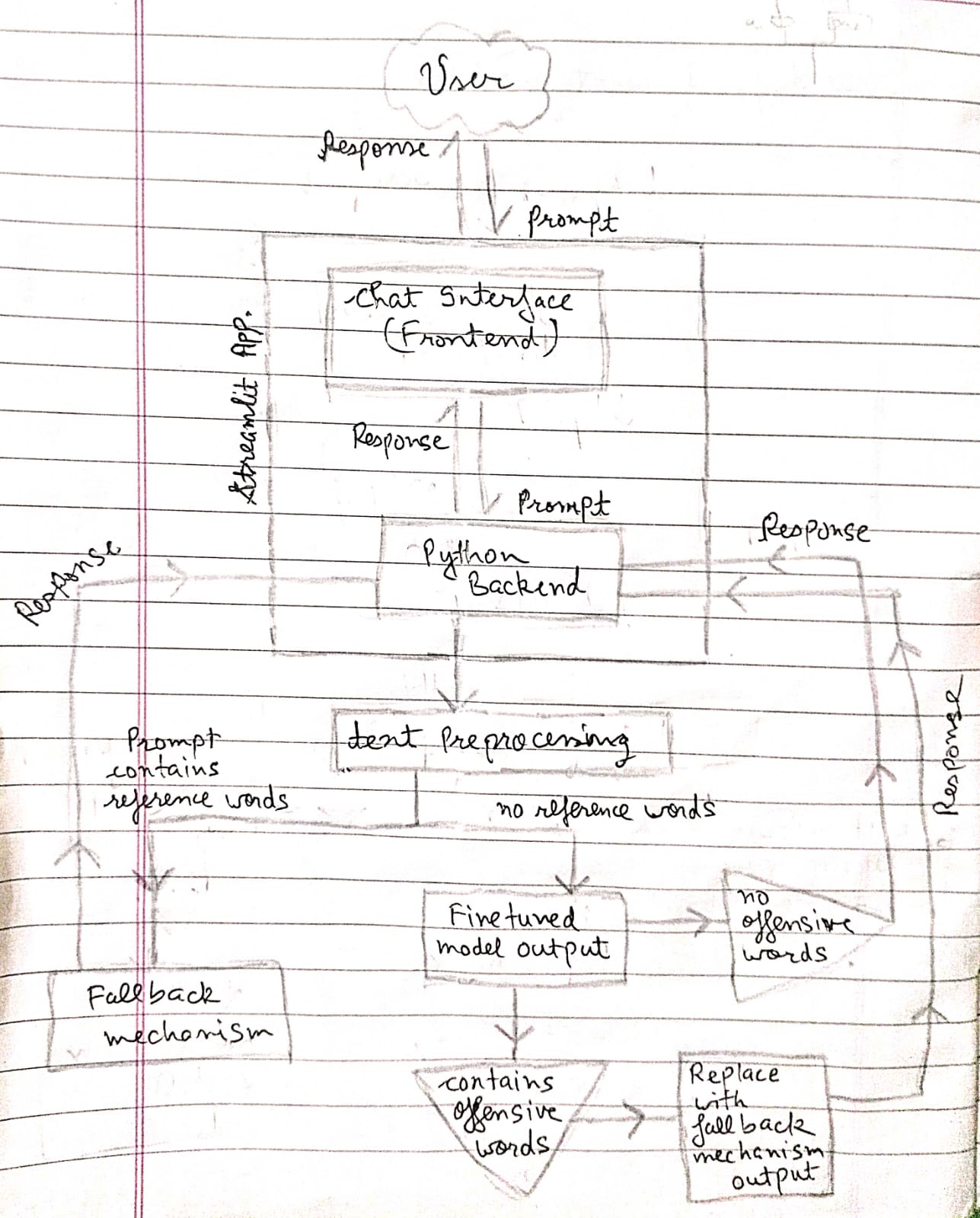
To assess model performance, the following metrics were used:

* **ROUGE (Recall-Oriented Understudy for Gisting Evaluation):** Evaluated the overlap between the generated text and reference text, focusing on recall.
* 

#### 

#### 

#### **3.6 Application UML Diagram**



#### 

#### 

#### 

#### 

#### **4. Recommendations and future scope/Improvements**

**6.1 Key Insights**

* **Model Performance:** T5 showed versatile performance even though it was only trained 1 epoch due to resource constraint.
* **LLama 2 -7b chat hf-** model is too large to finetune on local machine or google colab free tier resulting in halting of training mid-way due to GPU vram overflow.
* **Data Sensitivity:** The models were sensitive to data quality, highlighting the importance of thorough data cleaning and preprocessing, also different models have different requirements for training and inference data structures and format.
* **Model responses** - The bot can answer questions related to common phenomena and universal truths as they formed the majority in the quora dataset. But it stumbles while answering logical and subjective( directed towards it) questions due to the very same reason.

**6.2 Recommendations and future scope**

* **Data Augmentation:** To improve model robustness, consider using data augmentation techniques, such as paraphrasing and synonym replacement, to diversify training data.
* **Fine-tuning for Specific Tasks:** Further fine-tuning on domain-specific data can significantly enhance model performance in specialized areas. I could only run the fine-tuning for 1 epoch (cycle) due to resource constraint, while the recommended epochs for proper fine-tuning is around 5. If in future more computing power is available then the model performance can be greatly increased.
* **Hybrid Models:** Exploring hybrid models that combine the strengths of BERT's contextual understanding with GPT's generative capabilities could provide superior results. This can be done by allowing different models to com]pare/vote for the best output amongst themselves using some comparison mechanism or pass one model’s answer through another so it can be further improved with both model’s knowledge.
* **Retrain with varying hyperparameters** - One most common method is to use a variety of hyperparameter and observe the performance of the model, then choosing the best one. But again this is only possible with abundant computing power so that training the same model 5-6 times with different parameters can be done in parallel.

#### **7. Conclusion**

This project successfully explored the capabilities of various NLP models for text generation. Developing a state-of-the-art question-answering model leveraging the Quora Question Answer Dataset. I created an AI system capable of understanding and generating accurate responses to a variety of user queries, mimicking a human-like interaction.The findings suggest a path forward for improving model performance and practical applications, including deploying these models in real-world applications.