NLP- Individual Final Report

Name- Sairam Venkatachalam

Introduction:

The project we worked on, "Talk to a President" aimed to develop a chatbot capable of answering questions as if it were a historical leader, specifically drawing from the speeches of past U.S. and Russian presidents. The project leveraged state-of-the-art natural language processing techniques, including fine-tuning a pre-trained language model (GPT-2)m using larger Language models for inferencing, creating a vector database, and implementing Retrieval Augmented Generation for efficient document retrieval. The chatbot's responses were designed to emulate the tone, style, and substance of historical speeches, offering an interactive way to explore the rhetoric of world leaders.

The shared work involved multiple components: collecting and pre-processing speech data, fine-tuning the GPT-2 language model, integrating speeches from both nations, creating embeddings for document storage, and implementing the chatbot interface. My individual contributions focused on two critical areas: fine-tuning the GPT-2 model for U.S. presidential speeches and building the Chroma vector database for document retrieval. Although I did contribute in other areas too such as streamlit and presentation.

Description of Individual Work

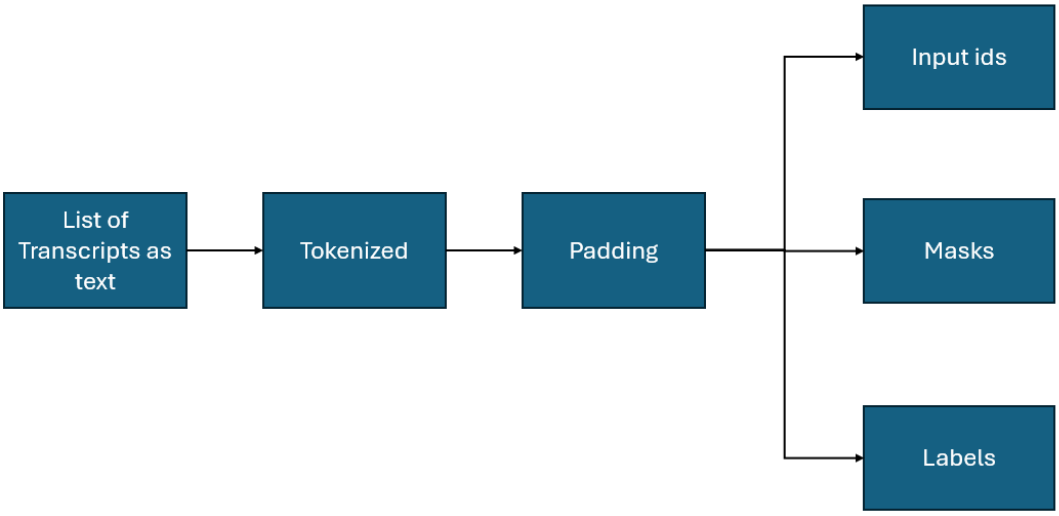
Finetuning the GPT-2 Model

We decided on 2 approaches to create language models that mimic presidents:

1. Option 1 was using a small Language model of less than 1B parameters, that could be trained on our dataset using our instances, and the computational resources we had available
2. Option 2 was using a much larger Language model of more than 10B parameters, that was smart enough to understand specific prompts, and used the context retrieved by RAG in order to answer the specific questions. These models would not be trained, but instead be used for inferencing

In order to train a model on our dataset, we had to select a small model that was open source. After testing a few models from hugging face like the falcon instruct and Qwen models, we decided to use GPT-2's base structure which contained 120 Million Parameters.

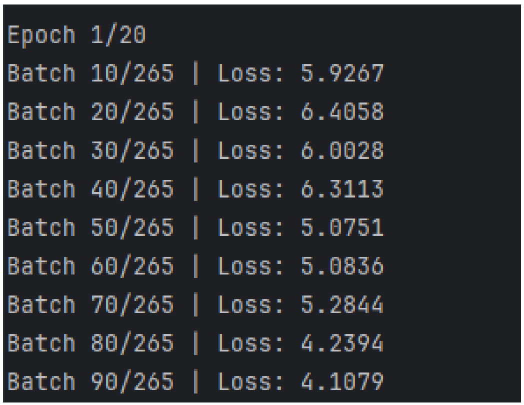
In order to finetune the GPT-2 model, U.S. and Russian presidential speeches were prepared and input to the model along with labels. This is shown below:



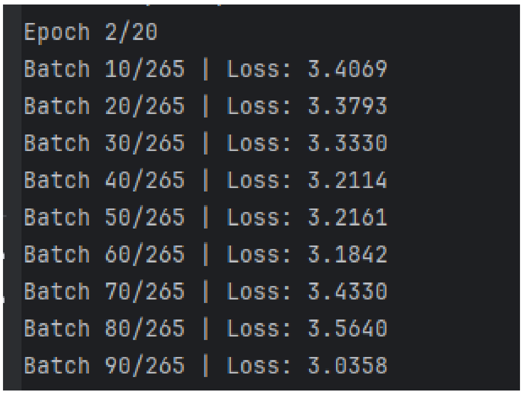
The figure above shows how the input data was prepared for training. The masks are basically 1's and 0's depending on whether it is a token or padded id. The attention masks ensured that padding tokens were ignored during training.

Training Process

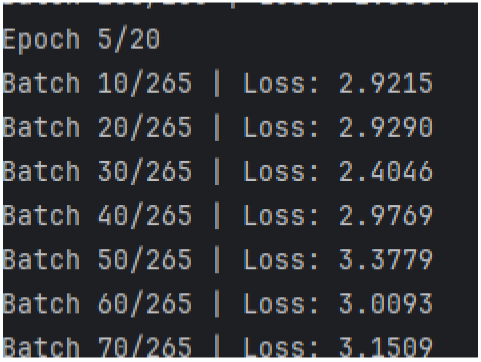
The fine-tuning process minimized the cross-entropy loss of the GPT-2 model over 20 epochs. The Adam optimizer with weight decay was employed to ensure smooth convergence. Training was conducted on a GPU, leveraging PyTorch for efficient computation. After fine-tuning on U.S. speeches, the model demonstrated improved ability to emulate the rhetorical style of U.S. presidents.



```
Epoch 1/20
Batch 10/265 | Loss: 5.9267
Batch 20/265 | Loss: 6.4058
Batch 30/265 | Loss: 6.0028
Batch 40/265 | Loss: 6.3113
Batch 50/265 | Loss: 5.0751
Batch 60/265 | Loss: 5.0836
Batch 70/265 | Loss: 5.2844
Batch 80/265 | Loss: 4.2394
Batch 90/265 | Loss: 4.1079
```



```
Epoch 2/20
Batch 10/265 | Loss: 3.4069
Batch 20/265 | Loss: 3.3793
Batch 30/265 | Loss: 3.3330
Batch 40/265 | Loss: 3.2114
Batch 50/265 | Loss: 3.2161
Batch 60/265 | Loss: 3.1842
Batch 70/265 | Loss: 3.4330
Batch 80/265 | Loss: 3.5640
Batch 90/265 | Loss: 3.0358
```



```
Epoch 5/20
Batch 10/265 | Loss: 2.9215
Batch 20/265 | Loss: 2.9290
Batch 30/265 | Loss: 2.4046
Batch 40/265 | Loss: 2.9769
Batch 50/265 | Loss: 3.3779
Batch 60/265 | Loss: 3.0093
Batch 70/265 | Loss: 3.1509
```

Creating the Chroma Vector Database

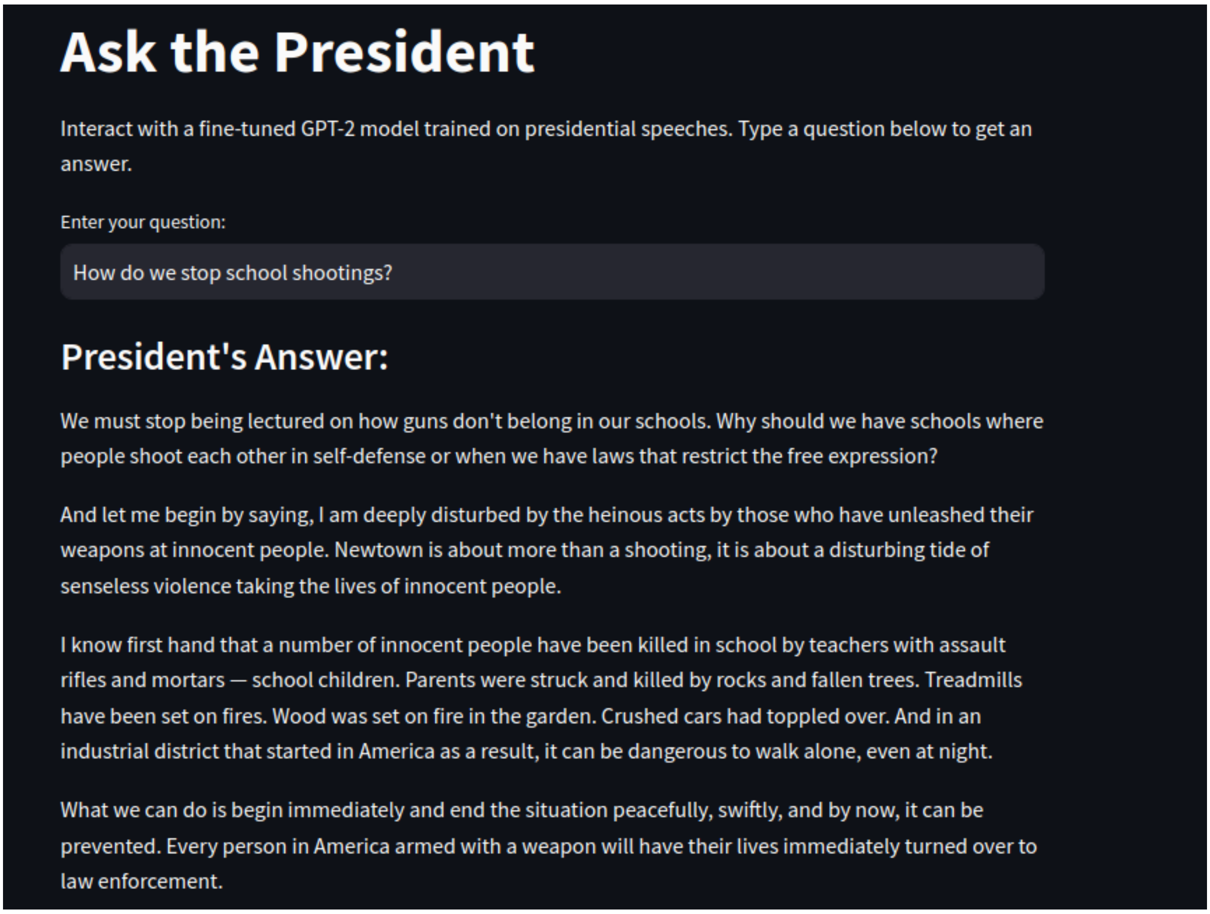
The Chroma vector database was a critical component for document retrieval in the chatbot. My work involved building the database using LangChain's recursive text splitter and embedding techniques.

Using LangChain's recursive text splitter, I divided lengthy speeches into manageable chunks, ensuring semantic coherence within each chunk. These chunks were processed for embedding generation.

Each text chunk was embedded into vector space using the sentence transformer embedding model. The embeddings were then stored in an in-memory vector database using Chroma, which enabled fast and accurate similarity-based retrieval.

Results

The fine-tuned GPT-2 model produced responses that effectively mirrored the tone and content of historical U.S. presidential speeches. For example:



Ask the President

Interact with a fine-tuned GPT-2 model trained on presidential speeches. Type a question below to get an answer.

Enter your question:

How do we stop school shootings?

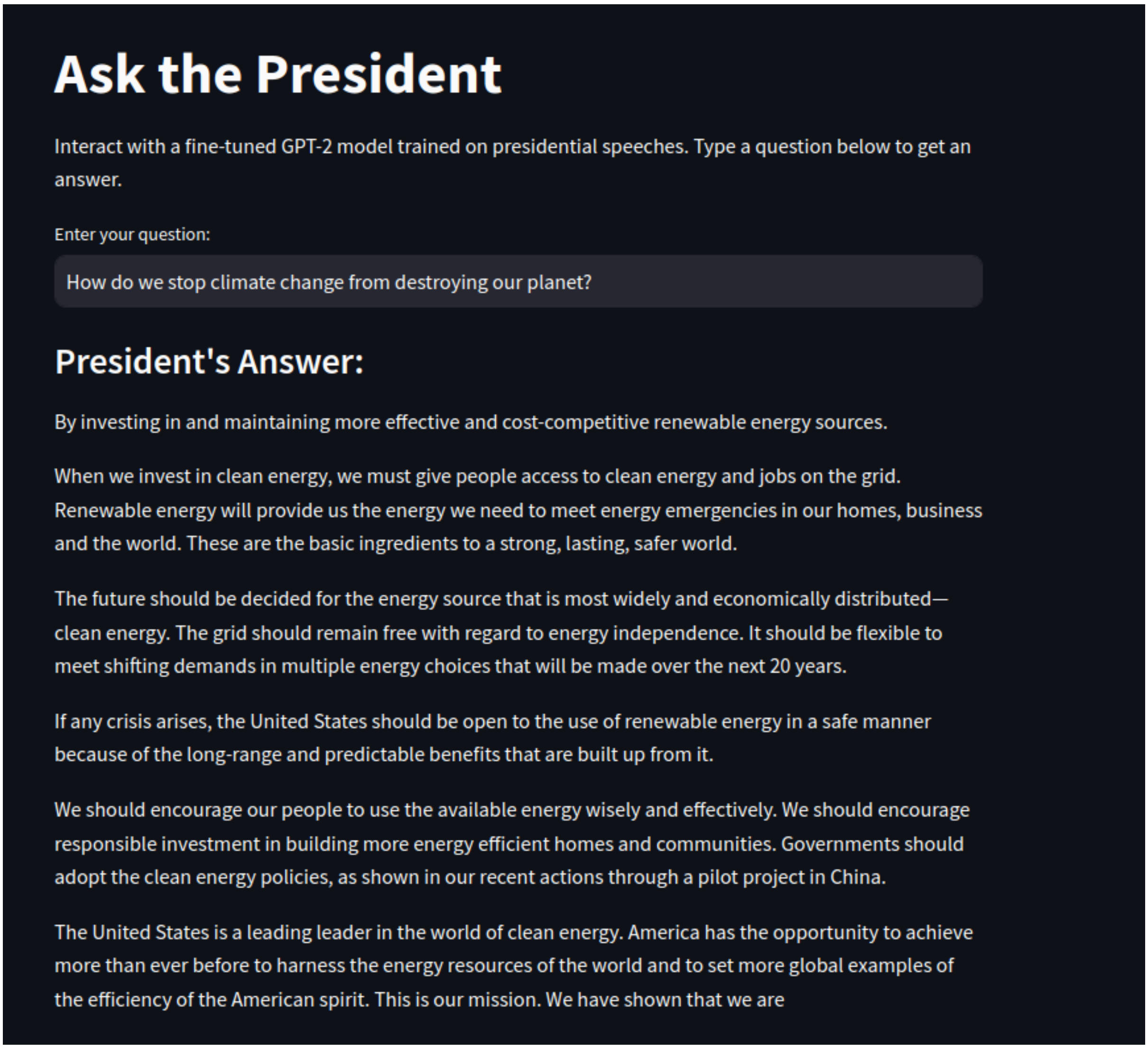
President's Answer:

We must stop being lectured on how guns don't belong in our schools. Why should we have schools where people shoot each other in self-defense or when we have laws that restrict the free expression?

And let me begin by saying, I am deeply disturbed by the heinous acts by those who have unleashed their weapons at innocent people. Newtown is about more than a shooting, it is about a disturbing tide of senseless violence taking the lives of innocent people.

I know first hand that a number of innocent people have been killed in school by teachers with assault rifles and mortars — school children. Parents were struck and killed by rocks and fallen trees. Treadmills have been set on fires. Wood was set on fire in the garden. Crushed cars had toppled over. And in an industrial district that started in America as a result, it can be dangerous to walk alone, even at night.

What we can do is begin immediately and end the situation peacefully, swiftly, and by now, it can be prevented. Every person in America armed with a weapon will have their lives immediately turned over to law enforcement.



Ask the President

Interact with a fine-tuned GPT-2 model trained on presidential speeches. Type a question below to get an answer.

Enter your question:

How do we stop climate change from destroying our planet?

President's Answer:

By investing in and maintaining more effective and cost-competitive renewable energy sources.

When we invest in clean energy, we must give people access to clean energy and jobs on the grid. Renewable energy will provide us the energy we need to meet energy emergencies in our homes, business and the world. These are the basic ingredients to a strong, lasting, safer world.

The future should be decided for the energy source that is most widely and economically distributed—clean energy. The grid should remain free with regard to energy independence. It should be flexible to meet shifting demands in multiple energy choices that will be made over the next 20 years.

If any crisis arises, the United States should be open to the use of renewable energy in a safe manner because of the long-range and predictable benefits that are built up from it.

We should encourage our people to use the available energy wisely and effectively. We should encourage responsible investment in building more energy efficient homes and communities. Governments should adopt the clean energy policies, as shown in our recent actions through a pilot project in China.

The United States is a leading leader in the world of clean energy. America has the opportunity to achieve more than ever before to harness the energy resources of the world and to set more global examples of the efficiency of the American spirit. This is our mission. We have shown that we are

Note: the above output was from an initial streamlit app. The UI of the final streamlit looks significantly different.

These outputs were further integrated into the final chatbot by my teammate Smit, who extended the fine-tuning to include speeches from Russian leaders.

Summary and Conclusions

My contributions to the "Talk to a President" project were essential in enabling the chatbot to emulate historical presidential rhetoric and retrieve relevant contextual information efficiently. The fine-tuning of GPT-2 provided a strong foundation for generating stylistically accurate responses, while the Chroma vector database ensured rapid and precise document retrieval which helped the larger models during inferencing.

Key Learnings

1. Fine-tuning pre-trained language models requires careful pre-processing and might need hyperparamter tuning

2. It is also worth noting that we may have overfit slightly on our data, as prompt engineering on the fine tuned model did not help much, suggesting that its understanding of language in general was overshadowed by presidential speech patterns
3. Efficient document retrieval enhances the contextual relevance of chatbot responses. This is quite generalizable, and can also help in anomaly detection because we can measure how close an input query is to the retrieved document

Future Improvements

1. Expanding the dataset to include more diverse speeches for broader coverage.
2. Optimizing the retrieval pipeline to handle larger-scale databases with reduced latency.
3. Exploring advanced fine-tuning techniques such as LoRA (Low-Rank Adaptation) for computational efficiency.

Code Origin Calculation

Most of the code I worked on were my own ideas, not copied from any github repository or internet code. I did however, use Generative AI like ChatGPT to assist in code writing , mostly to do with explaining errors, and setting up functions for data preparation or wrapper functions.

References

1. Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). "Attention Is All You Need." arXiv preprint arXiv:1706.03762.

3. Hugging Face Transformers Documentation. https://huggingface.co/docs/transformers

4. LangChain Documentation. https://langchain.readthedocs.io

5. https://www.researchgate.net/figure/GPT-2-model-architecture-The-GPT-2-model-contains-N-Transformer-decoder-blocks-as-shown_fig1_373352176

6. OpenAI. (2024). *ChatGPT* [Large language model]. https://chatgpt.com