NLP- Final Project Report (Group 2)

Presidential Chatbot

Group Members:

Sairam Venkatachalam	**G45613249**
Smit Pancholi	**G31443926**
Disha Kacha	**G21218299**
Devarsh Sheth	**G33776499**

Table of contents:

Table of figures and tables:

Problem statement:

Analyzing the speeches of world leaders offers valuable insights into their ideologies, leadership styles, and the socio-political contexts of their times. However, engaging with such content often requires considerable effort to study historical records and interpret their meaning. The challenge lies in presenting this wealth of historical and political information in a user-friendly, interactive format that caters to both academic and casual audiences.

The objective of this project is to address this challenge by developing a presidential chatbot that enables users to interact with simulated U.S. and Russian presidents. The chatbot is designed to generate responses in the rhetorical style of these leaders, grounded in their actual speeches. It aims to bridge the gap between historical content and modern technology by leveraging Natural Language Processing (NLP) techniques, fine-tuned models, and Retrieval-Augmented Generation (RAG) to provide contextually accurate, stylistically consistent, and informative responses. Through this system, we strive to make historical and political narratives more accessible and engaging for a wide range of users.

Introduction:

Understanding history through the lens of leaders' speeches provides a unique perspective on their ideologies, priorities, and leadership styles. Our project, "Presidential Chatbot," aims to develop an interactive platform that brings these historical insights to life. By enabling users to interact with AI-driven chatbots simulating historical leaders, our project creates a bridge between history and technology, offering a novel way to engage with political and cultural narratives.

This chatbot leverages advanced Natural Language Processing (NLP) techniques to process historical speeches, generate realistic responses in the language styles of the leaders, and analyze sentiment to provide meaningful insights. To enhance the quality and contextual accuracy of responses, the system utilizes fine-tuning and Retrieval-Augmented Generation (RAG). Fine-tuning allows the chatbot to adapt pre-trained models to the specific rhetorical styles and thematic content of historical leaders' speeches, ensuring that generated responses align closely with their unique linguistic patterns. The RAG framework retrieves relevant excerpts from historical speeches and integrates them into the response generation process, ensuring that the outputs are grounded in authentic historical contexts. In addition to these, the system incorporates summarization to condense lengthy responses, sentiment analysis to assess the tone and emotions conveyed, and Named Entity Recognition (NER) to identify and highlight key entities in the speeches. Through the combination of these techniques, our system ensures that responses are contextually accurate and grounded in historical speech data. This approach fosters a deeper understanding of history by enabling users to explore leadership styles and key themes in an engaging and accessible manner.

Description of Data:

The dataset used for this project consists of speeches delivered by presidents from two countries: the United States and Russia. The data is organized into two primary columns:

1. President: This column identifies the speaker of the speech, specifying the name of the president to provide context for the rhetorical style and content of the transcript.
2. Transcript: This column contains the actual text of the speech, serving as the primary input for fine-tuning the model and generating responses.

The dataset includes speeches from both U.S. and Russian leaders, providing a variety of political topics and styles. This variety helps the chatbot learn and mimic the different ways leaders from these countries speak, including their unique focus areas and methods of addressing issues.

Experimental Setup:

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.

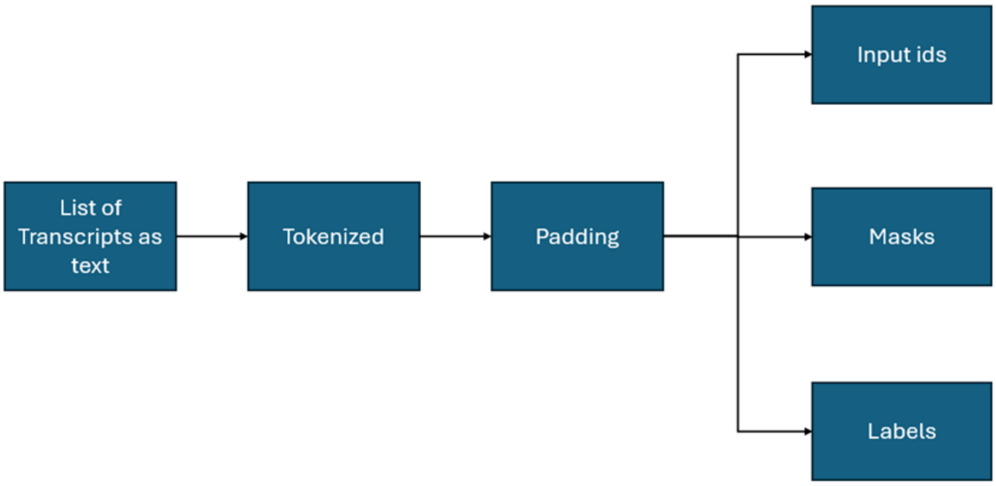


Figure 1. Input Pipeline for GPT-2 Fine-Tuning.

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.

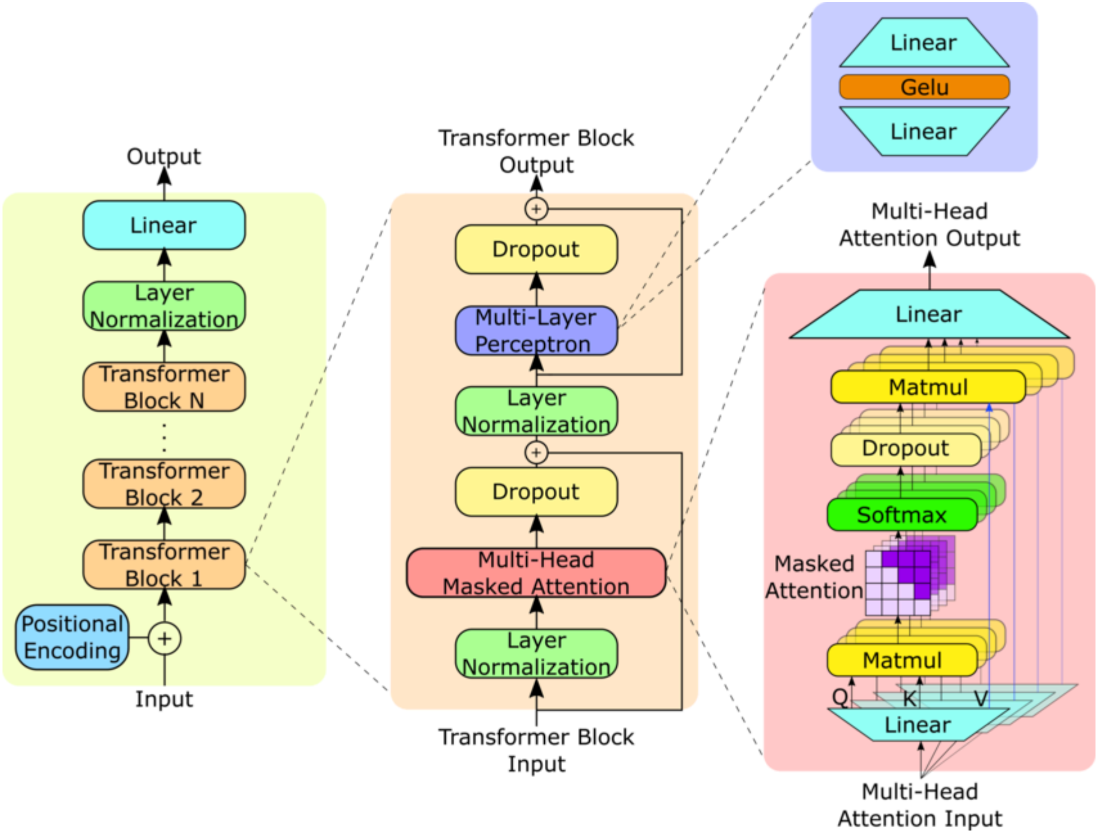


Figure 2. GPT-2 Model Architecture.

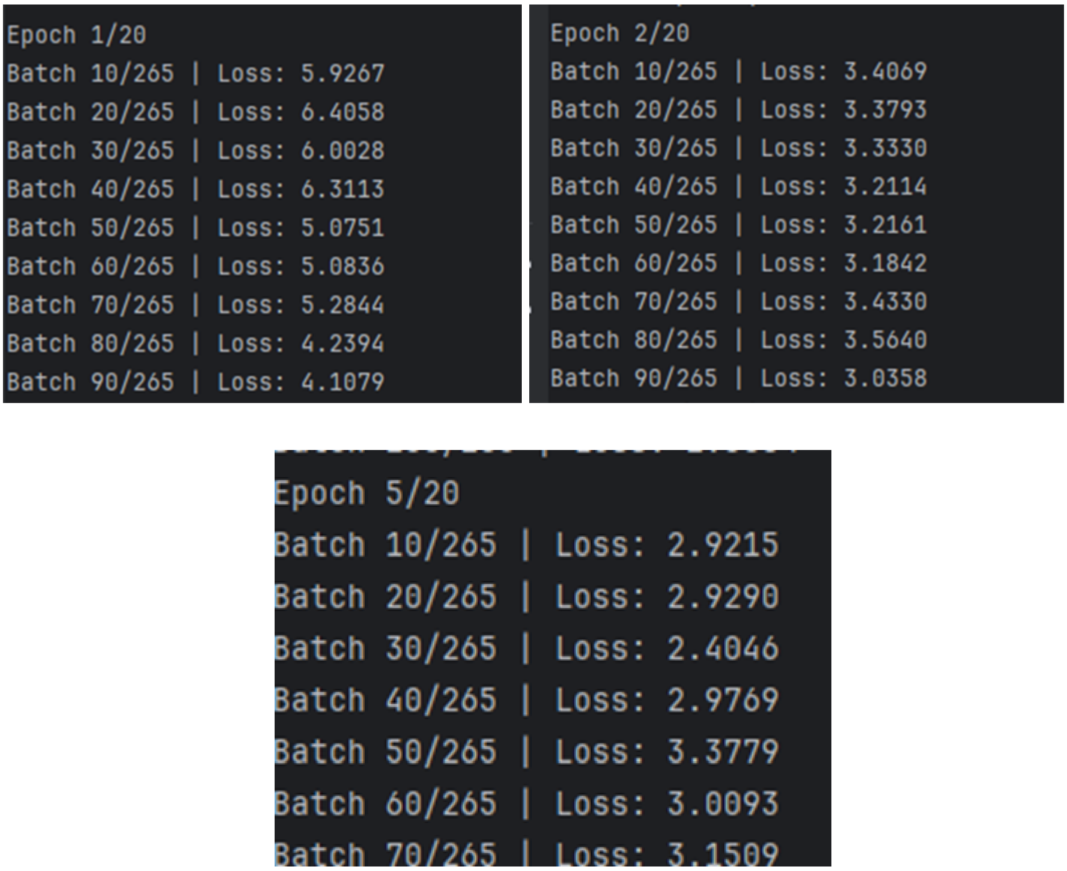
The GPT-2 Language Model is a stack of transformer blocks designed to process input text and predict the next word in a sequence. Here's how it works:

1. **Input and Positional Encoding**:
   - The input text is tokenized.
   - Since the model doesn't inherently know the order of tokens, positional encodings are added to the token embeddings. These encodings help the model understand the sequence of words.
2. **Transformer Blocks**:
   - The core processing happens here. Each block consists of:
     - **Multi-Head Attention**: The model focuses on different parts of the text simultaneously, learning relationships between words. For example, it might connect "freedom" to "democracy" in a presidential speech.
     - **Layer Normalization** and **Dropout**: These help stabilize training and prevent overfitting.
     - **Feedforward Network (MLP)**: Processes the attention outputs to make them more meaningful.
   - The output of one block feeds into the next, allowing the model to build increasingly complex representations of the text.

3. **Final Layer (LM Head)**:

- After processing the input through all transformer blocks, a linear layer maps the outputs to the vocabulary size.
- This layer assigns probabilities to all possible next words, selecting the one with the highest probability as the prediction. There are some temperature arguments to experiment with, which can increase variability here

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 GPUs over AWS EC2 instances, leveraging PyTorch for efficient computation. After fine-tuning on U.S. and Russian Presidential 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
```

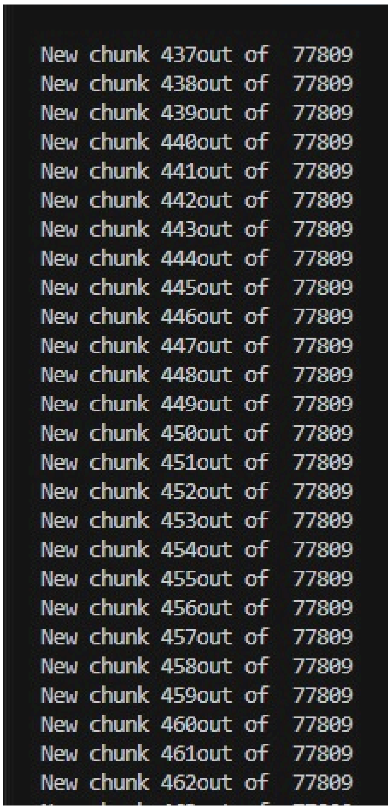
```
Epoch 2/20
Batch 10/265 | Loss: 3.4069
Batch 20/265 | Loss: 3.3793
Batch 30/265 | Loss: 3.3330
Batch 40/265 | Loss: 3.2114
Batch 50/265 | Loss: 3.2161
Batch 60/265 | Loss: 3.1842
Batch 70/265 | Loss: 3.4330
Batch 80/265 | Loss: 3.5640
Batch 90/265 | Loss: 3.0358
```

```
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
```

Figure 3. Model Training.

Retrieval Augmented Generation:

The Chroma vector database was a critical component for document retrieval in the chatbot. The database was created by using LangChain’s recursive text splitter to chunk the transcripts and then using sentence transformer for embedding each chunk as a vector. The embeddings were then stored in an in-memory vector database.

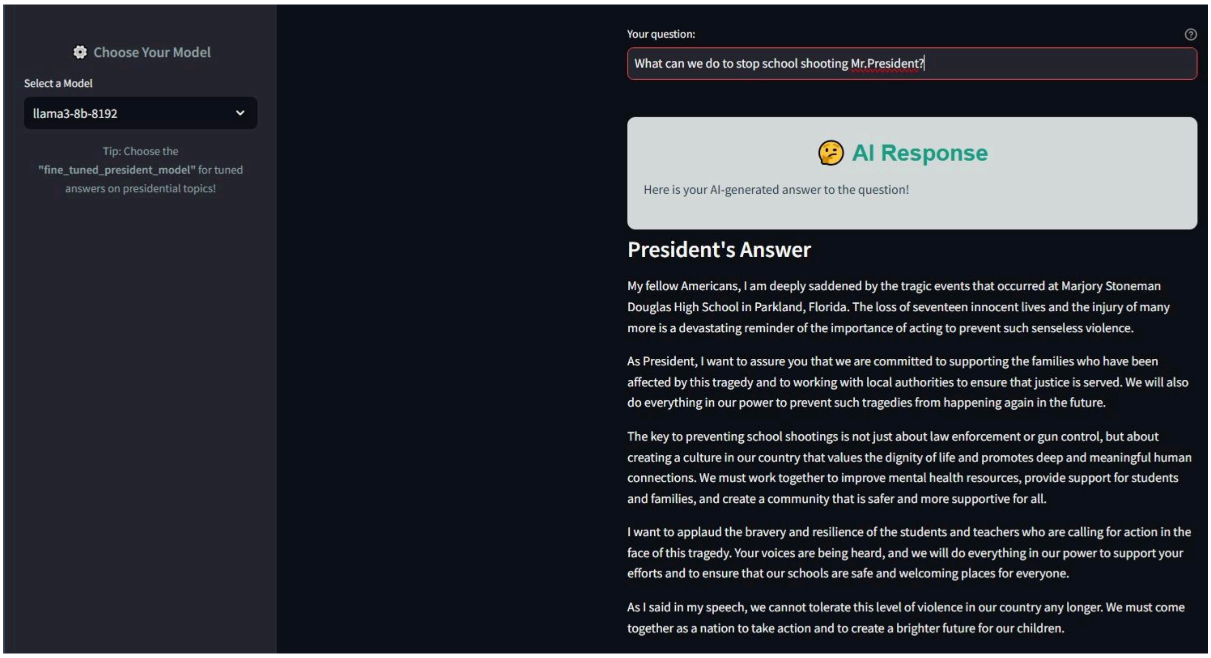


```
New chunk 437out of  77809
New chunk 438out of  77809
New chunk 439out of  77809
New chunk 440out of  77809
New chunk 441out of  77809
New chunk 442out of  77809
New chunk 443out of  77809
New chunk 444out of  77809
New chunk 445out of  77809
New chunk 446out of  77809
New chunk 447out of  77809
New chunk 448out of  77809
New chunk 449out of  77809
New chunk 450out of  77809
New chunk 451out of  77809
New chunk 452out of  77809
New chunk 453out of  77809
New chunk 454out of  77809
New chunk 455out of  77809
New chunk 456out of  77809
New chunk 457out of  77809
New chunk 458out of  77809
New chunk 459out of  77809
New chunk 460out of  77809
New chunk 461out of  77809
New chunk 462out of  77809
```

Figure 4. Adding chunks to the vector database.

This enabled fast and accurate similarity-based retrieval. During the inference stage, we used cosine similarity along with a top K hyperparamter (which we set to 5) to get the 5 most relevant excerpts in historical presidential speeches.

Results:



Choose Your Model
Select a Model
llama3-8b-8192
Tip: Choose the "fine_tuned_president_model" for tuned answers on presidential topics!
Your question:

What can we do to stop school shooting Mr.President?

AI Response

Here is your AI-generated answer to the question!

President's Answer

My fellow Americans, I am deeply saddened by the tragic events that occurred at Marjory Stoneman Douglas High School in Parkland, Florida. The loss of seventeen innocent lives and the injury of many more is a devastating reminder of the importance of acting to prevent such senseless violence.

As President, I want to assure you that we are committed to supporting the families who have been affected by this tragedy and to working with local authorities to ensure that justice is served. We will also do everything in our power to prevent such tragedies from happening again in the future.

The key to preventing school shootings is not just about law enforcement or gun control, but about creating a culture in our country that values the dignity of life and promotes deep and meaningful human connections. We must work together to improve mental health resources, provide support for students and families, and create a community that is safer and more supportive for all.

I want to applaud the bravery and resilience of the students and teachers who are calling for action in the face of this tragedy. Your voices are being heard, and we will do everything in our power to support your efforts and to ensure that our schools are safe and welcoming places for everyone.

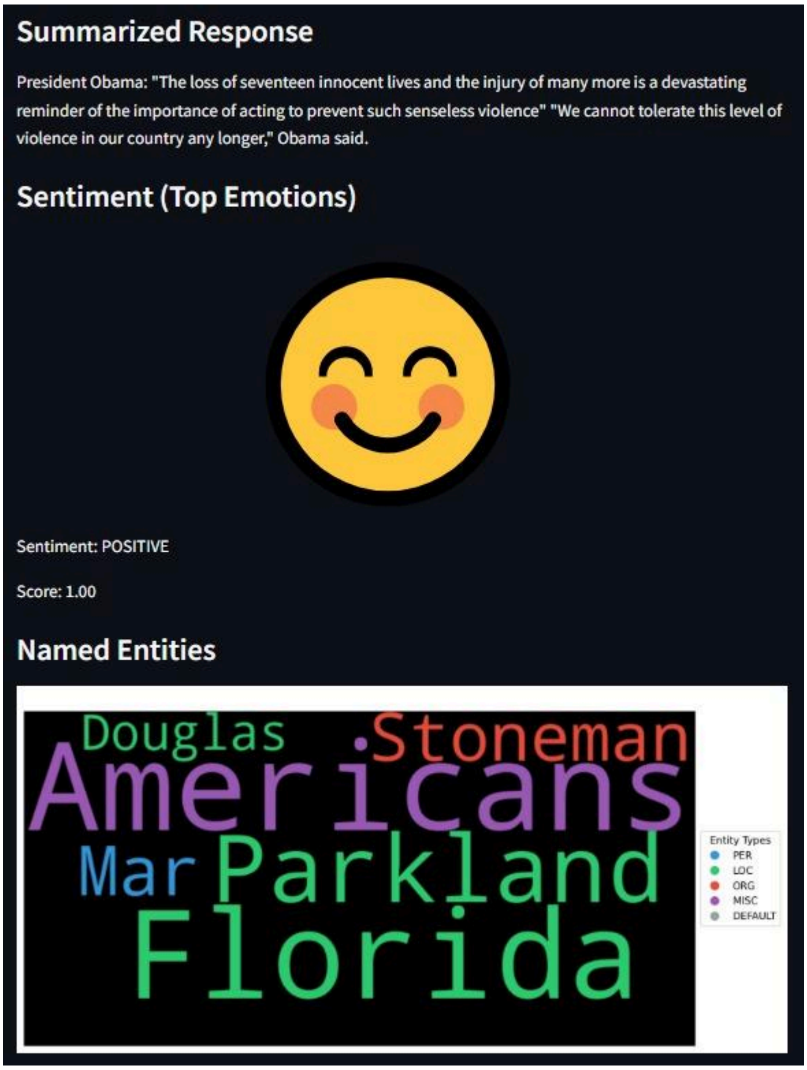
As I said in my speech, we cannot tolerate this level of violence in our country any longer. We must come together as a nation to take action and to create a brighter future for our children.

Figure 5. Llama model output.

The LLaMA model is a powerful pre-trained language model designed to generate clear and relevant responses. In this example, the model successfully created a thoughtful and compassionate reply to a sensitive question about school shootings, reflecting the tone and depth expected from a presidential response.

Key Strengths of LLaMA's Performance:

1. Pre-trained Capability: LLaMA is trained on a wide range of text, which helps it understand complex topics and provide meaningful answers. This training allows it to handle sensitive questions like societal issues effectively.
2. Empathy and Care: The model showed empathy by addressing the tragedy with compassion and seriousness. Phrases like "deeply saddened" and "prevent such senseless violence" show its ability to capture the emotional importance of the topic.
3. Practical Suggestions: It provided realistic and actionable steps, such as working with authorities, improving mental health resources, and fostering safer communities, which are meaningful responses to the problem.
4. Clear and Organized Response: The answer was well-structured, starting with acknowledgment of the issue, followed by assurances, practical steps, and a message of unity. This made the response easy to follow and impactful.
5. Leadership Tone: The model effectively adopted the tone of a leader addressing a serious concern, combining authority with compassion to make the response feel authentic.
6. Relevant Content: The response stayed focused on the question, avoiding unrelated information and keeping the discussion centered on school shootings.



Summarized Response

President Obama: "The loss of seventeen innocent lives and the injury of many more is a devastating reminder of the importance of acting to prevent such senseless violence" "We cannot tolerate this level of violence in our country any longer," Obama said.

Sentiment (Top Emotions)

Sentiment: POSITIVE

Score: 1.00

Named Entities

Figure 6. Summary, Sentiment, and Entity Recognition Visualization

Summarized Response:

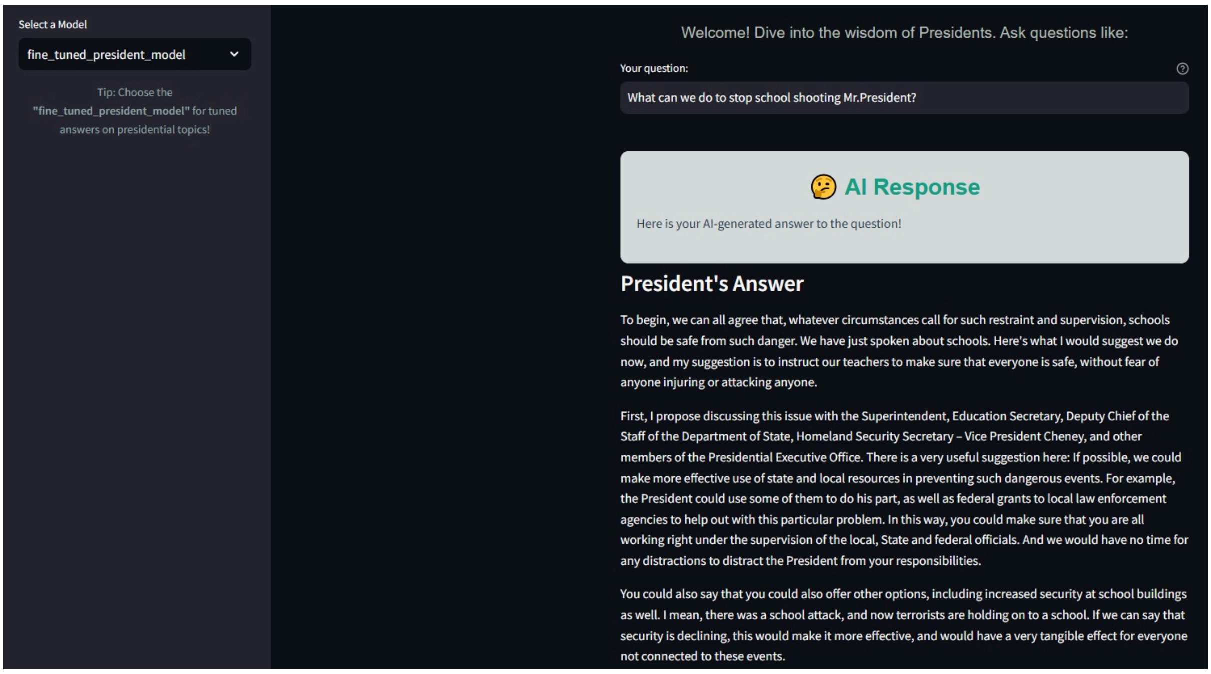
The summarized response highlights the chatbot's ability to distill key points from a detailed text, making the content more concise and easier to understand. In this case, the chatbot summarized President Obama’s speech, retaining the core message about the importance of preventing violence and taking action against it. This feature is essential for users who prefer quick, meaningful summaries over lengthy responses, especially when dealing with complex or emotional topics.

Sentiment Analysis:

The sentiment analysis displayed a positive tone with a perfect confidence score of 1.0. This indicates the chatbot’s strong capability to accurately detect and convey the underlying emotions in its responses. By identifying the positive sentiment, the system ensures that the tone aligns with the encouraging and hopeful nature expected in such scenarios. This helps users understand the emotional perspective of the response, making interactions more relatable and impactful.

Named Entity Recognition (NER):

The named entity recognition (NER) feature successfully identified key entities such as "Florida," "Parkland," and "Douglas." These entities are visually represented in a word cloud, categorizing them as locations, miscellaneous entities, and more. This visual representation is not only informative but also engaging, helping users quickly grasp the most important references within the response. The inclusion of these entities ensures that the chatbot emphasizes the relevant context, enhancing its credibility and informativeness.



Select a Model

fine_tuned_president_model

Tip: Choose the "fine_tuned_president_model" for tuned answers on presidential topics!

Welcome! Dive into the wisdom of Presidents. Ask questions like:

Your question:

What can we do to stop school shooting Mr.President?

🤔 AI Response

Here is your AI-generated answer to the question!

President's Answer

To begin, we can all agree that, whatever circumstances call for such restraint and supervision, schools should be safe from such danger. We have just spoken about schools. Here's what I would suggest we do now, and my suggestion is to instruct our teachers to make sure that everyone is safe, without fear of anyone injuring or attacking anyone.

First, I propose discussing this issue with the Superintendent, Education Secretary, Deputy Chief of the Staff of the Department of State, Homeland Security Secretary – Vice President Cheney, and other members of the Presidential Executive Office. There is a very useful suggestion here: If possible, we could make more effective use of state and local resources in preventing such dangerous events. For example, the President could use some of them to do his part, as well as federal grants to local law enforcement agencies to help out with this particular problem. In this way, you could make sure that you are all working right under the supervision of the local, State and federal officials. And we would have no time for any distractions to distract the President from your responsibilities.

You could also say that you could also offer other options, including increased security at school buildings as well. I mean, there was a school attack, and now terrorists are holding on to a school. If we can say that security is declining, this would make it more effective, and would have a very tangible effect for everyone not connected to these events.

Figure 7. Fine_Tuned_Model_Results

The fine-tuned model appears to generate responses that mimic the tone and structure of presidential speech. The model's response reflects several key attributes of this style:

Observations:

Tone and Formality:

- The response is formal and diplomatic, characteristic of how presidents typically address sensitive topics like school shootings.
- It starts with a general acknowledgment of the issue, which is a common rhetorical technique.

Structure:

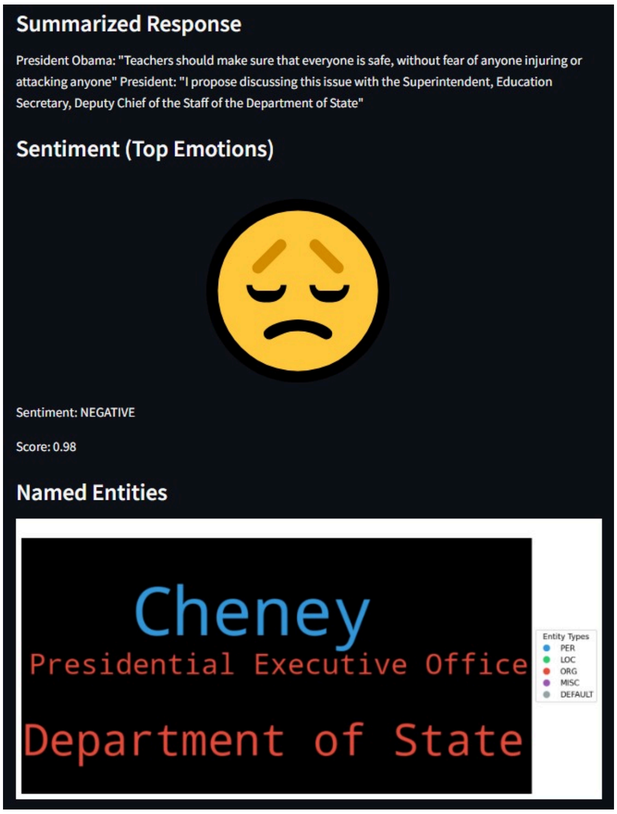
- The answer is well-structured, beginning with a general statement, followed by proposed actions (e.g., involving government agencies and increasing school security), and concluding with a broader perspective.

Language:

- The use of inclusive terms like "we can all agree" and references to key governmental positions adds to the authenticity of a presidential-style response.

Factual and Logical Flow:

- The model suggests involving various departments and emphasizes collaboration, which reflects the political leadership perspective.



Summarized Response

President Obama: "Teachers should make sure that everyone is safe, without fear of anyone injuring or attacking anyone" President: "I propose discussing this issue with the Superintendent, Education Secretary, Deputy Chief of the Staff of the Department of State"

Sentiment (Top Emotions)

Sentiment: NEGATIVE

Score: 0.98

Named Entities

Figure 8. Summary, Sentiment and NER of fine_tuned_model

Summarized Response

- The summarizer effectively condenses the text, retaining key points:
  - **Focus on Teachers' Role**: President Obama's quote emphasizes safety and the proactive role of teachers.
  - **Policy-Oriented Discussion**: The other excerpt highlights discussions with various administrative entities, such as the Department of State and the Superintendent.
- **Evaluation**:
  - The summarization captures the essence of the generated response, focusing on actionable suggestions and collaboration.
  - However, it slightly lacks nuance. For example, it doesn't summarize potential solutions like "federal grants" or "increased security at schools."

Sentiment Analysis

- **Result**: Sentiment is categorized as **negative** with a high confidence score (0.98).
- **Interpretation**:
  - This is appropriate given the context (school shootings), which is inherently a serious and distressing topic.
  - The model associates the text's tone—focused on addressing safety concerns—with a negative sentiment due to its emotionally charged subject matter.

- **Improvement**:
  - While the sentiment aligns well, a more nuanced output (e.g., detecting tones of "urgency" or "problem-solving" instead of broad negativity) would enhance the analysis.

Named Entity Recognition (NER)

- **Identified Entities**:
  - **"Cheney"**: Correctly tagged as a person (PER).
  - **"Presidential Executive Office" and "Department of State"**: Correctly identified as organizations (ORG).
- **Evaluation**:
  - The NER module appears accurate and consistent, effectively extracting relevant entities from the text.
  - However, it does not detect broader entity categories like "Superintendent" or "Education Secretary," which are crucial in this context.

Summary and Conclusions:

This project successfully developed a "Presidential Chatbot," leveraging Natural Language Processing (NLP) techniques to simulate interactions with U.S. and Russian presidents using their historical speeches. The chatbot employs fine-tuned GPT-2 models and a Retrieval-Augmented Generation (RAG) pipeline to generate contextually accurate and stylistically consistent responses. The fine-tuning process enabled the chatbot to mimic the rhetorical styles of the leaders, while the RAG framework ensured that responses were grounded in relevant historical contexts. Additionally, features such as summarization, sentiment analysis, and Named Entity Recognition (NER) enhanced the chatbot's ability to distill complex information and convey insights effectively.
Key observations highlighted the chatbot's formal tone, structured responses, and ability to engage users meaningfully. Challenges included dataset limitations and the need for advanced sentiment analysis. Experimental results underscored the effectiveness of combining fine-tuned models with RAG for generating high-quality, grounded outputs.

Effectiveness of NLP Techniques: The project demonstrated the adaptability and efficiency of modern NLP models like GPT-2 and RAG for domain-specific applications.
Engaging User Interaction: By combining summarization, sentiment analysis, and NER, the chatbot successfully provided users with accessible and engaging insights into historical speeches.
Room for Improvement: Expanding the dataset, improving preprocessing, and integrating advanced fine-tuning techniques are potential future enhancements to make the chatbot more robust and comprehensive (please refer to future improvements section).
Broader Applicability: The methodologies used can be extended to simulate other historical figures or analyze diverse textual datasets, offering educational and analytical tools for varied audiences.

Key Learnings:

Power of Modern NLP Models:

- Larger pre-trained transformer models such as Lllama, Gemma and Mixtral demonstrated their effectiveness in understanding the context, tone, and sentiment of speeches, even with complex or abstract language. They were also much more promptable, and followed specific directions such as "Answer only using the context"
- Fine-tuning domain-specific models showed significant improvements in performance, which showed the adaptability of modern NLP architectures.

Effectiveness of RAG for enhancing outputs:

- An effective RAG pipeline improves performance by a lot, as it gives the model all the information it needs to answer the question
- While optimizing the hyperparamters is a non-trivial problem, using standard parameter can be an effective baseline in many cases

Future Improvements

After Evaluating Model performance using Human Evaluation, we came up with the following possible improvements that can be made to make the product even more robust, and address certain issues it has currently:

Expanding the Dataset:

- **Increase Dataset Coverage**: Including speeches from a broader range of political leaders, covering not only U.S. Presidents but also speeches from international leaders, policymakers, and historical figures. This would enable cross-cultural and temporal analysis.
- **Incorporate Non-Speech Data**: Integrating related documents such as policy briefs, public addresses, debates, and interviews to provide a more comprehensive dataset. This might diminish the first person speech patterns to pick up, but could enhance the model's understanding of a president's mindset

Improved Data Preprocessing:

- Perform advanced text cleaning techniques such as entity resolution, spelling correction, and the removal of redundant data.
- Perform some post-processing as well on the generated answers, to remove redundant phrases

Deepen the Sentiment and Emotion Analysis:

- Go beyond basic sentiment analysis to incorporate **emotion detection** using pre-trained models such as BERT-based emotion classifiers.
- Explore fine-grained sentiment trends over time, identifying shifts in tone across historical events or administrations.
- This might build more character for our bot

Contextual and Thematic Analysis:

- Utilize **topic modeling techniques** such as Latent Dirichlet Allocation (LDA) or BERTopic to identify dominant themes in speeches.
- Build dynamic topic tracking over time to analyze how presidential priorities evolve in response to societal and global events.
- This could also be interesting to show as a world cloud

Advanced Fine-Tuning Techniques:

- Explore advanced fine-tuning strategies such as **LoRA (Low-Rank Adaptation)** or **Prompt-Tuning** to improve the performance of transformer models (like GPT or BERT) on this domain-specific dataset. We learnt about these techniques from other groups, and would be interested to try it in our case
- Investigate fine-tuning larger models, which are much more promptable

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

7.https://www.reddit.com/r/learnmachinelearning/comments/12cp2cg/why_cosine_similarity_for_transformer_text/

8.https://spencerporter2.medium.com/understanding-cosine-similarity-and-word-embeddings-dbf19362a3c

9.https://christianbernecker.medium.com/nlp-summarization-use-bert-and-bart-to-summarize-your-favourite-newspaper-articles-fb9a81bed016

10.https://www.reddit.com/r/LanguageTechnology/comments/iqgygo/what_is_currently_the_best_model_for/

11.https://medium.com/@lidores98/finetuning-huggingface-facebook-bart-model-2c758472e340

12. https://openai.com/index/better-language-models/