LLM as Presidential Candidate: Analyzing Shifts in Political Rhetoric Over Time

Soline Boussard, Niki Ekstrom, Michelle Hewson, Sydney Levy
Massachusetts Institute of Technology

Abstract

This study explores the evolution of sentiment, rhetoric, and political issue focus in U.S. presidential and vice-presidential debates from 1960 to 2024, leveraging large language models (LLMs) to uncover patterns and insights. Sentiment analysis of the debates reveals a downward trend, with recent debates exhibiting more negative rhetoric than earlier debates. Fine-tuning the GPT-4o mini model on subsets of the debate data, divided by Republican and Democratic across 2 time periods, created 4 distinct LLMs - each representing a simulated candidate in a presidential debate. The debate responses from each fine-tuned model reveal a more limited topic range and reduced overlap in party priorities over time. Temporal context strongly influences debate themes, with the simulated modern debates focusing on economic issues, while the earlier simulated debates address a broader array of societal concerns. These findings highlight the growing polarization and evolving focus of American political rhetoric.

1 Introduction

Presidential debates offer a unique lens through which the priorities, rhetoric, and political landscape of the United States can be examined. These debates not only reflect societal concerns of their time, but also play a crucial role in shaping public opinion and electoral outcome. For instance, the most recent U.S. presidential debate in 2024 attracted over 67 million viewers (Grynbaum, 2024), highlighting the significant influence these debates have on elections and public discourse. By analyzing shifts in tone, focus, and sentiment, political debates provide valuable insights into the broader evolution of the nation's political and social landscape (Hobolt et al., 2024).

However, traditional methods of analyzing debate rhetoric often struggle to capture nuanced trends in the sentiment, rhetoric, and issue prioritization over time (Chandra and Saini, 2021; Üveges and Ring, 2023; Taubenfeld et al., 2024). Recent advancements in LLMs have created new opportunities for understanding and analyzing political language. When fine-tuned on domain-specific data, these models can reveal trends such as sentiment shifts, rhetorical patterns, and changes in topic prioritization among political parties.

This study leverages these advancements to examine the evolution of rhetoric in U.S. presidential debates from 1960 to 2024. Sentiment analysis is applied to debate transcripts to assess variations in the emotional tone of candidates' rhetoric—positive, negative, or neutral—across parties and election cycles. To complement this, 4 LLMs are fine-tuned on debate transcripts and prompted with standardized questions to generate simulated candidate responses. These simulated responses provide insights into topics and rhetorical patterns that may not be explicitly addressed in the debates. By applying traditional NLP techniques, such as sentiment analysis and topic modeling, to these generated responses, the study uncovers deeper trends in candidate rhetoric, including shifts in tone and topic prioritization.

2 Related Works

2.1 Sentiment Analysis Methodologies

A significant body of research has focused on sentiment analysis in the political domain, primarily leveraging data from social media (Severyn and Moschitti, 2015; Gupta et al., 2020) and news outlets (Hossain et al., 2021). These studies often aim to understand public opinion, predict election outcomes, or analyze political rhetoric. However, much of this work has emphasized platforms where the sentiment of everyday citizens is expressed rather than focusing on the candidates' language itself. This research narrows this scope to an im-

pactful yet less explored area: U.S. presidential and vice-presidential debates.

Understanding political debates requires a nuanced analysis due to the complexity of persuasive rhetoric, which basic models often fail to capture. Valence Aware Dictionary and Sentiment Reasoner (VADER), a rule-based model designed for social media, provides a simple baseline for observing sentiment shifts but lacks contextual sensitivity (Hutto and Gilbert, 2014). Addressing these limitations, Bidirectional Encoder Representations from Transformers (BERT) offers bidirectional context interpretation, enabling it to capture nuanced sentiment in political speech, as demonstrated in analyses of the 2020 U.S. election (Chandra and Saini, 2021). RoBERTa, an advanced framework built on BERT, improves performance by leveraging larger datasets, refined hyperparameters, and dynamic masking, making it particularly effective for analyzing complex political sentiment shifts (Liu et al., 2019).

Zheng et al. (2023) provide a foundational approach for using "LLM-as-a-judge" frameworks to assess subjective human preferences, achieving over 80% alignment with human evaluations on benchmarks like MT-Bench and Chatbot Arena (Zheng et al., 2023). Using a frontier model to score political sentiment in U.S. presidential debate transcripts allows for the capture of nuanced shifts in tone and rhetoric across different parties and time periods. This approach adapts the "LLM-as-a-judge" methodology to the political domain, offering a scalable way to assess sentiment in complex, persuasive language. Thus, GPT-4o mini will be used as a judge since it is a cost-efficient model developed by OpenAI that uses its advanced contextual understanding to perform an analysis of political debates across time periods and contexts (OpenAI, 2024). To analyze the sentiment of the debates between 1960-2024, VADER, RoBERTa, and GPT-4o mini are all used.

2.2 Fine-Tuning LLMs to Generate Political Rhetoric

Beyond using LLMs to perform traditional sentiment analysis, fine-tuning LLMs offers a powerful approach for generating text based on specific topics and historical contexts (Li et al., 2024). By leveraging architectures that have the ability to understand varying language patterns and contexts, these frontier LLMs can be trained on domain-specific datasets to capture underlying themes. Fine-tuning LLMs on subsetted datasets from particular time periods and political parties will allow one to generate text that reflects an era's specific language and rhetoric (Taubenfeld et al., 2024). Park et al. (2023) describes a method for an LLM to simulate an identity by feeding a model different prompts (Park et al., 2023). This research adopts this approach to simulate presidential candidates, as it will provide valuable insights into how political rhetoric may shift across different historical and partisan contexts, thus offering a unique tool for exploring changes in political discourse over time.

2.3 LLM Potential and Limitations for Political Analysis

Recent research on applying LLMs to political discourse analysis has further highlighted their abilities and limitations. Liu et al. (2024) suggest that LLMs are an important complement to human analysis for evaluating debate performances because LLMs allow for rapid analysis across multiple debates, making them far more scalable (Liu et al., 2024). Churina & Jaidka (2024) emphasize the importance of high-quality datasets and prompting strategies to improve the quality of political speech generation, especially in polarized contexts (Churina and Jaidka, 2024).

Taubenfield et al. (2024) focuses on fine-tuning GPT-models to simulate historical political rhetoric in debates of political partisans. Their work emphasizes the ability of models to capture historical patterns of discourse and find that LLMs often exhibit inherent social biases that can worsen their ability to simulate realistic political debates (Taubenfeld et al., 2024). Our study builds on this foundation by incorporating sentiment analysis and topic tracking across different time periods.

3 Data

The American Presidency Project provides both presidential and vice presidential debate transcripts from 1960 to 2024 (The American Presidency Project, n.d.). Each transcript is tagged with metadata including the year, debate type (presidential, vice presidential, or primary), political party association, and participant names.

4 Methodology

4.1 Data Preprocessing

To prepare the data for analysis, the candidates' responses were divided into chunks corresponding

to each question posed by the moderator. Additionally, text from the moderators was excluded to ensure that the models focus exclusively on the rhetoric and language of the candidates.

4.2 Sentiment Analysis of Political Rhetoric Over Time

Multiple sentiment analysis methodologies were used to analyze the rhetoric within presidential and vice presidential debates between 1960 and 2024. Candidate responses were divided into 250-character chunks to allow for manageable processing. Sentiment scores were then calculated to identify the emotional tones within the candidates' rhetoric.

The initial baseline sentiment analysis was performed using VADER, a lexicon-based model. Next, a fine-tuned RoBERTa model, pre-trained on 58 million tweets, was applied to assess sentiment with a deeper contextual understanding. Finally, GPT-4o mini was prompted (prompt found in Appendix C) to interpret the sentiment of the debate fragments, offering a more nuanced sentiment analysis.

4.3 Simulating Presidential Candidates: LLM Fine-tuning on Debate Data Subsets

To further understand the contrast between the political rhetoric of today and of the past, 4 LLMs were fine-tuned to represent simulated political candidates from the Democratic and Republican parties and 2 historical time periods to help uncover contrasts that might not be apparent through the sentiment analysis alone. The debate transcripts were subsetted by both parties and by time period for debates between 1976-1984 and 2016-2024. These 2 time periods were chosen due to their strikingly different sentiment trends. By fine-tuning LLMs on these subsetted transcripts, the produced models can replicate the speech patterns of political candidates in a specific time period and party.

To fine-tune the GPT-4o mini model, the data needed to be converted into question-answer pairs. In debates it is common for candidates to interrupt each other or not directly answer the question posed by the moderator, thus the raw debate text did not contain 1:1 correspondence between questions asked by the moderator and candidate responses. To reformat, the moderator questions were removed from the transcripts and the data was subset by political party and time period. GPT-4o mini was then prompted with each candidate response and asked to produce a question that would reasonably insight the provided response (prompt found in Appendix D). GPT-4o mini was then fine-tuned on each of these synthetic question-response pairs (still divided by party and time period), with each model trained for 3 epochs using a batch size of 1 (training details found in Appendix E.

After fine-tuning, a set of 14 unique prompts was provided to the fine-tuned models to simulate debate questions across a range of topics in each time period (prompts found in Appendix G). The prompts were crafted to focus on topics that were likely to surface in a debate during both time periods. For example, *"How will you create jobs and reduce unemployment across the nation?"* Since the models had a temperature of 0.7, they were prompted with each question 10 times to account for model variation in responses.

4.4 Analyzing Debate Simulations from Fine-Tuned Candidate Models

To analyze and compare the outputs of the fine-tuned models, several computational techniques were employed:

4.4.1 Sentiment Analysis

Sentiment analysis was implemented through RoBERTa and GPT-4o mini, as outlined in Section 4.2 methodology, to extract the sentiment trends from the simulated debate responses.

4.4.2 Term Frequency-Inverse Document Frequency (TF-IDF)

TF-IDF analysis was used to identify significant terms and phrases (unigrams, bigrams, and trigrams) in the generated responses. This analysis focused on comparing the relative importance of words across different time periods and political parties.

Before performing TF-IDF analysis, the generated responses were preprocessed by converting text to lowercase, removing non-alphabetic characters, and filtering out common stopwords, including frequent but uninformative words such as "uh" and "im." Lemmatization was then applied to reduce words to their base forms. The top 10 terms or phrases for each group were identified based on their TF-IDF scores and visualized using word clouds.

4.4.3 Latent Dirichlet Allocation (LDA)

LDA analysis was used to better understand the key topics that were focused on in the simulated debates

across the fine-tuned models. For this topic modeling analysis, responses were preprocessed similarly to the TF-IDF analysis. Text was converted to lowercase, non-alphabetic characters were removed, and stopwords (including domain-specific ones) were filtered out. Lemmatization ensured linguistic consistency.

The combined dataset of all 4 model-generated responses to 14 questions was vectorized using the CountVectorizer module from scikit-learn, configured with the custom stopword list developed during preprocessing. The result was a document-term matrix representing the frequency of words across the corpus. An LDA model was fit to this matrix, specifying 14 topics to align with the 14 questions posed to the models and fixed random seed to ensure reproducibility. This analysis provided a structured representation of underlying themes in the model-generated responses, facilitating a uniform evaluation of the topics that surfaced in debate speech across years and political parties.

4.4.4 Cosine Similarity

To evaluate the semantic similarity of model-generated responses, each response from the fine-tuned models was vectorized using the "all-MiniLM-L6-v2" model from SentenceTransformers. Cosine similarity was calculated to measure semantic alignment between responses, focusing on within-model consistency and cross-model similarity.

Within-model consistency was assessed by grouping responses by the model (based on time period and party) as well as the specific question asked. Cosine similarities were calculated for all response vectors, and the average similarity for each question (within the same model) was computed. These question-level averages were aggregated to determine the overall consistency of each model.

To assess similarity across models, cosine similarities were computed between responses generated by different models for the same question. For each pair of models, the average cosine similarity across all responses to a given question was calculated and stored in a similarity matrix. Diagonal entries of the matrix represented within-model similarity, using the values for consistency across repeated responses. This similarity matrix served as a quantitative summary of the semantic alignment between models and was later visualized to explore patterns of similarity across time periods and political parties.

5 Results

5.1 Sentiment Analysis of Political Rhetoric Over Time

The VADER model (Figure 1) shows that sentiment has had a slight downward trend between 1960 and today, however it classifies the majority of the transcripts as positive, or having sentiment scores above 0.

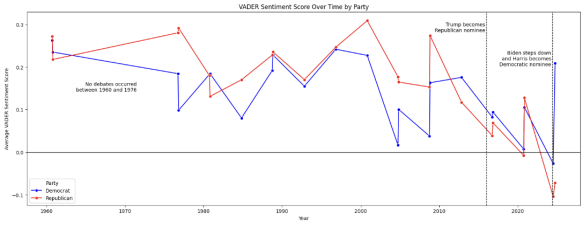

Figure 1: Sentiment Analysis using VADER

In contrast, the RoBERTa model (Figure 2) consistently classifies the debate transcripts as negative (below 0).

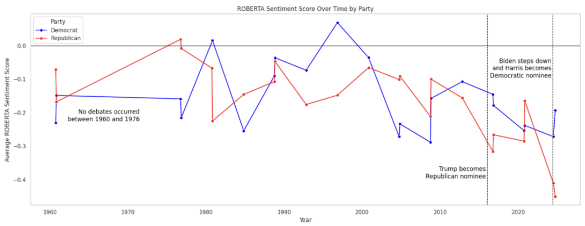

Figure 2: Sentiment Analysis using RoBERTa

GPT-4o mini (Figure 3) produced average sentiment scores fluctuating around 0, indicating more balanced sentiment throughout the debates.

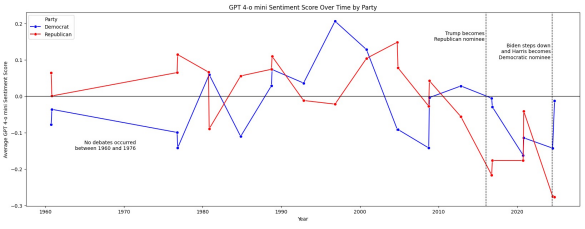

Figure 3: GPT-4o mini Sentiment Analysis

Notably, all three models show agreement that the Republican party has seen an increase in negativity since 2010, coinciding with the candidacy of Former President Trump. Meanwhile, the sentiment within the Democratic party increased positively following President Biden's resignation from the candidacy and Vice President Harris's nomination. Additionally, when comparing sentiment be-

tween 1976-1984 vs. today we can see that overall sentiment has become more negative within both parties.

5.2 LLM as Presidential Candidate: Simulating Debates with Fine-Tuned LLMs

5.2.1 Sentiment of Fine-Tuned Models

Unlike with the sentiment analysis on actual debate transcripts, analysis of the simulated debates suggests sentiment is more positive today (2016-2024) compared to the past (1976-1984) (Figure 4). The simulated Democratic responses have become much more positive, while Republican responses only increase slightly over time. They also indicate that there is a greater disparity in sentiment between the generated rhetoric from parties today compared to in the past.

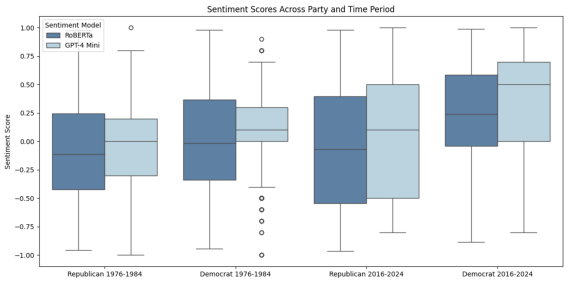

Figure 4: Sentiment Analysis on Fine-Tuned Models' Generated Text

5.2.2 TF-IDF Analysis

The TF-IDF analysis of the 1976–1984 and 2016–2024 time periods highlights differences in the generated model responses for Republicans and Democrats (Figures 5 and 6). In the 2016–2024 responses, only 2 of the top-10 most common phrases were the same for the 2 parties, with one of these being non-political in nature ("youre going"). Democratic responses from this period show a stronger emphasis on specific political issues, such as "environmental sustainability" and "equal opportunity," while the Republican responses tend to emphasize national identity and patriotic themes, as reflected in phrases like "make America great" and "American people."

In contrast, the 1976-1984 period shows greater overlap between the 2 parties, with terms like "civil right" and "federal government" appearing in both Republican and Democratic responses. However, party-specific themes still emerge, with Republicans focusing on individuals (e.g., "Mr. Mondale," "Mr. Carter") and Democrats highlighting systemic issues such as "affirmative action" and "health care."

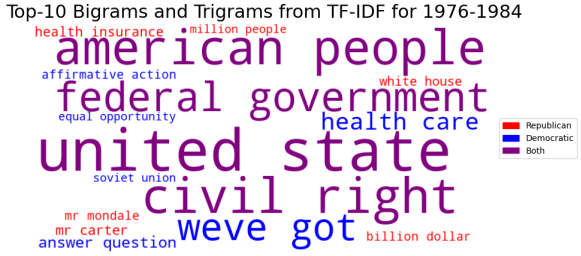

Figure 5: TF-IDF Word Cloud from 1976-1984 Model's Generated Text

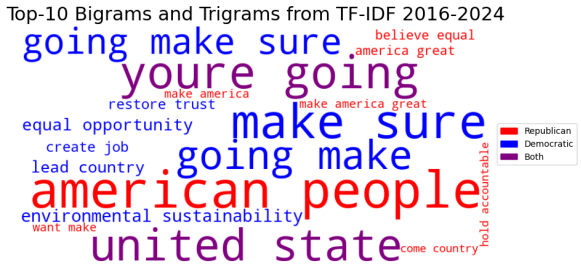

Figure 6: TF-IDF Word Cloud from 2016-2024 Model's Generated Text

5.2.3 LDA Analysis

The LDA analysis identified 14 distinct topics across the model-generated responses, reflecting key themes such as healthcare, education, civil rights, environmental policy, and the economy. The full topic list and top 10 keywords for each topic are listed in Appendix H.

Topic distributions across the simulated debates reveal notable differences in the key issues emphasized across time periods, while party affiliation appears to have a less influence on topic similarity. Debates within the same time period, such as Democratic and Republican debates from 2016-2024, show high topic similarity with a cosine similarity score of 0.97 (Figure 7). In contrast, debates separated by decades, such as Democratic debates from 2016-2024 and 1976-1984, exhibit lower similarity scores (0.60), indicating substantial shifts in discourse over time.

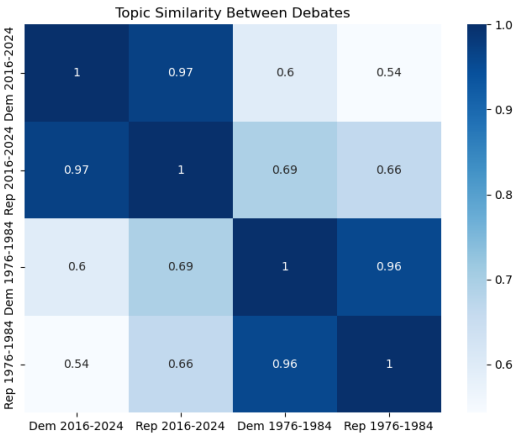

Figure 7: Topic Similarity on Fine-tuned Models

The heatmap of topic similarity (Figure 8) highlights these temporal shifts. Debates from 1976-1984 feature a broader range of topics, including "Presidential Leadership" (0.23) and "Civil Rights Issues" (0.12), while debates from 2016-2024 were more focused on "Economic Opportunities" (0.30 and 0.37 for Republican and Democratic debates, respectively).

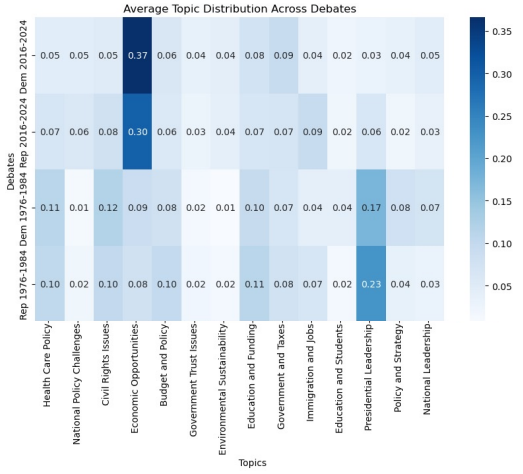

Figure 8: LDA Topic Distribution on Fine-tuned Models' Generated Text

5.2.4 Cosine Similarity

Cosine similarity analysis (Figure 9) measured both the similarity between different models and the consistency of each model's responses to the same question across multiple iterations.

Cross-model cosine similarity analysis reveals that the most similar models are Republicans and Democrats from 1976-1984 (cosine similarity of 0.37), suggesting that the largest differences in debate rhetoric occur across time periods rather than between political parties. The least similar models are Democrats from 2016-2024 compared to Republicans from 1976-1984 (0.22), as these models span both time period and party.

The average within-model cosine similarity ranges from 0.33 to 0.44, indicating moderate variation in responses to the same question, likely due to the fine-tuning temperature of 0.7. The Republican model from 1976-1984 exhibits the highest internal consistency, while the Republican model from 2016-2024 shows the least consistency.

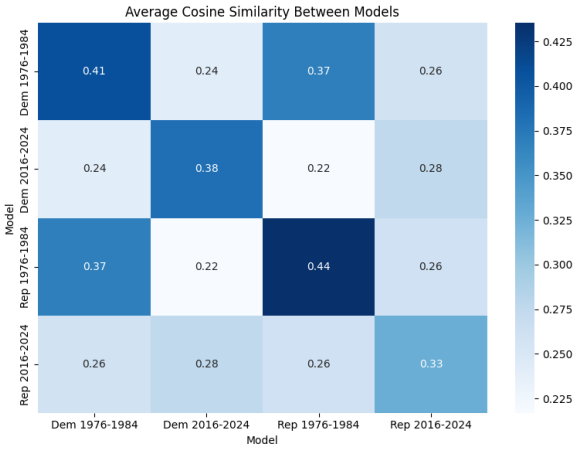

Figure 9: Cosine Similarity on Fine-tuned Models' Generated Text

6 Discussion

6.1 Temporal Shifts in Debate Rhetoric

The initial sentiment analysis findings suggest that the overall tone of U.S. presidential debates has fluctuated over time with a distinct downward trend. Each of the 3 models—VADER, RoBERTa, and GPT-4o mini—consistently indicated that the average sentiment of these debates has grown increasingly negative over time. Despite agreeing on the overall negative trend, the models show differences in baseline classifications: VADER predominantly classifies the debates as having a positive tone, RoBERTa labels most as negative, and GPT-4o mini produces scores mostly around 0, indicating a neutral sentiment.

One explanation for why RoBERTa classifies the majority of transcripts as negative while VADER classifies them as positive is because RoBERTa's improved contextual understanding allows it to capture more subtle or implied negative tones typical in political debates. On the other hand, VADER, being lexicon-based, might interpret isolated positive words more straightforwardly, leading to a generally positive sentiment. RoBERTa's sensitiv-

ity to complex, adversarial language aligns with the nuanced, often critical tone of debates, while VADER's simpler approach seems to miss these subtleties. GPT-4o mini's sentiment analysis results are less extreme than VADER and RoBERTa, indicated by its tendency to classify statements as neutral. Since GPT-4o mini is a language model trained on vast amounts of data sources while RoBERTa is trained on Tweets alone, GPT-4o mini has a broader context for analyzing sentiment, which may allow it to capture different trends, thus resulting in sentiment classifications centered around neutrality.

The debate speech generated from the LLMs fine-tuned as presidential candidates did not reflect the same negative sentiment trends. The Republican models showed similar sentiment distributions, both with medians around 0, but with the 1976-1984 Republican model having less variation in its responses (Figure 4). The Democratic models showed less agreement, with the 2016-2024 Democratic model having much higher median values for overall sentiment. This suggests that the fine-tuned LLMs could not replicate the same downward trend in overall tone observed in the actual debates. This could be attributed to the fact that the fine-tuned LLMs were generating responses to prompted questions and not engaging in a real debate against another candidate. As a result, the generated responses lacked the adversarial tone and frequency of attacks typical in modern debates.

6.2 Temporal Dynamics and Polarization in Political Debates

The findings of the LDA analysis, cosine similarity, and TF-IDF on the generated text from the simulated candidates reveal that time period has a more significant influence on the topics apparent in debates than party affiliation. These temporal shifts also reveal less diverse topics addressed in recent years. The disproportionate focus on Economic Opportunities in the 2016-2024 debates, with little attention to other issues, may signal increasing polarization in the political climate. Conversely, the broader range of topics in the 1976-1984 generated responses suggests a less polarized political environment, where candidates felt more comfortable to address a wider array of concerns.

The TF-IDF analysis further highlights these shifts in rhetorical focus. The results highlighted in Figures 5 and 6 suggest increasing polarization, with each party's model focusing on fewer general themes and less overall agreement.

The cosine similarity results additionally emphasize that political rhetoric has become increasingly polarized over time. Historically, there was more alignment between parties, as shown by the higher similarity between Republicans and Democrats from 1976–1984. This suggests that the ideological differences between time periods is more significant than the differences between parties.

7 Limitations and Future Work

This study provides insights into the changing political landscape between 1960 and today, but several opportunities for improvement remain. First, future research could include a more comprehensive analysis of additional time periods, rather than focusing solely on the 1976–1984 and 2016–2024 eras. By examining a broader range of election cycles, one could uncover more detailed trends in sentiment and rhetorical shifts, allowing for a richer understanding of how political language has evolved over decades.

Additionally, when fine-tuning LLMs, adjusting the model's temperature to a lower value could enhance consistency in responses. This would reduce variability when the same model is asked identical questions multiple times, thereby increasing cosine similarity within a model and yielding more reliable comparisons. These adjustments would help refine the methodology and create a more robust framework for analyzing political discourse over time.

To more accurately simulate debates between the fine-tuned LLMs, future work could involve feeding one model's generated text to the other model to replicate true debate dynamics.

Future work could also incorporate broader and larger datasets, such as additional congressional records, rallies, or campaign advertisements. Expanding the diversity of data sources would allow for a more comprehensive analysis of political rhetoric, providing insights that could be extended beyond presidential debates. This broader scope would enhance the generalizability of the findings and a more thorough understanding of how political language has evolved across different contexts, mediums, and topics.

References

Rohitash Chandra and Ritij Saini. 2021. Biden vs trump: Modeling us general elections using bert language model. *IEEE Access*, 9:128494–128505.

Svetlana Churina and Kokil Jaidka. 2024. Fine-tuning llms with noisy data for political argument generation. *Preprint*, arXiv:2411.16813.

M. Grynbaum. 2024. 67.1 million people watched harris and trump, outdrawing last debate. *The New York Times*.

Shloak Gupta, Sarah Bolden, Jay Kachhadia, A Korsunska, and J Stromer-Galley. 2020. Polibert: Classifying political social media messages with bert. In *Social, cultural and behavioral modeling (SBP-BRIMS 2020) conference. Washington, DC*.

Hobolt, Sara B., Katherina Lawall, and James Tilley. 2024. The polarizing effect of partisan echo chambers. *American Political Science Review*.

Arafat Hossain, Md. Karimuzzaman, Md. Moyazzem Hossain, and Azizur Rahman. 2021. Text mining and sentiment analysis of newspaper headlines. *Information*, 12(10).

C. Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):216–225.

Junyi Li, Tianyi Tang, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024. Pre-trained language models for text generation: A survey. *ACM Comput. Surv.*, 56(9).

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Zhengliang Liu, Yiwei Li, Oleksandra Zolotarevych, Rongwei Yang, and Tianming Liu. 2024. Llm-potus score: A framework of analyzing presidential debates with large language models. *Preprint*, arXiv:2409.08147.

OpenAI. 2024. GPT-4o mini: advancing cost-efficient intelligence. `https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/`. [Accessed 07-12-2024].

Joon Sung Park, Jamie O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, Michael S. Bernstein, Sean Follmer, Juho Han, Jürgen Steimle, and Nathalie Henry Riche. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22.

Aliaksei Severyn and Alessandro Moschitti. 2015. Twitter sentiment analysis with deep convolutional neural networks. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, page 959–962, New York, NY, USA. Association for Computing Machinery.

Amir Taubenfeld, Yaniv Dover, Roi Reichart, and Ariel Goldstein. 2024. Systematic biases in llm simulations of debates. *Preprint*, arXiv:2402.04049.

The American Presidency Project. n.d. Presidential candidates debates (1960-2024). Accessed: 2024-10-14.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

István Üveges and Orsolya Ring. 2023. Hunembert: A fine-tuned bert-model for classifying sentiment and emotion in political communication. *IEEE Access*, 11:60267–60278.

A Impact Statement

The intersection of politics and modern technology in today's society is a challenging road to navigate. With the prevalence of political echo chambers, which are becoming increasingly polarized due to curated news and social media feeds, it is critical to address the biases present in technologies, specifically LLMs. To further replicate the type of rhetoric that candidates from 1976-1984 and 2016-2024 produce, we trained 4 different LLMs on subsetted data. The subsetted data encompassed all presidential and vice-presidential debates in that time frame. Thus, these LLMs are susceptible to projecting the inherent biases that may exist in those subsetted datasets. For example, if the models are trained on data that contains false or exaggerated statements, the model may spread inaccurate information. Furthermore, it is well-known that LLMs are susceptible to "hallucinating," or generating untrue content. In the context of political discourse, hallucinations could possibly reinforce political biases, thus further polarizing public opinion.

Possible societal ramifications of this project are extensive, especially as AI technologies such as LLMs become more integrated in everyday life. One concern is using our fine-tuned LLMs as a news source, or as a ground truth. Although these models were trained to replicate a candidate's speech from a specific time period, the text they generate does not necessarily reflect exactly what a candidate would or would not say. If deployed in the real world without adequate safeguards, these models could contribute to reinforcing existing biases and contributing to a more polarized public. These biases may reinforce echo chambers and limit an individual's exposure to content and rhetoric outside of their political bubble.

Another key ethical concern arises from training model's on political figures' rhetoric such as Donald Trump. As a prominent figure in modern politics, Donald Trump has a unique pattern and style in his speeches. Training his speeches in AI systems can raise questions about intellectual property, fair use, and the potential for misrepresentation. Trump's words, although public, are his intellectual property, and training an AI model on his rhetoric could be viewed as exploiting his speech without adequate compensation or consent. Additionally, if an LLM is trained on Trump's speech and is able to mimic his tone, there is a risk of misrepresentation. People who encounter content generated by model might mistakenly believe it was written by Trump himself. This could blur the lines between his actual statements and AI-generated content, leading to confusion about the source and intent behind the words, potentially undermining trust in the authenticity of political discourse.

Thus, as LLMs continue to advance and play a greater role in today's society, it is imperative to implement safeguards, ethical guidelines, and transparency measures to mitigate these risks for their responsible deployment in society.

B Supplemental Information - Code and Data Used for Project

Visit our Google Drive with Data & Code to see the code and data used for this project.

C Prompt Used to Generate GPT-4o mini Sentiment Analysis Scores

"Please analyze the sentiment of the following text. Provide a score between -1 and 1, where -1 represents very negative sentiment, 0 represents neutral sentiment, and 1 represents very positive sentiment. Only give the score, no explanation. Here is the text: {text}"

D Prompt Used to Generate GPT-4o mini Moderator Questions

"Craft a question a debate moderator might ask that would naturally lead to the following response in a presidential debate. The question should be neutral, engaging, and directly related to the content of the provided response. Only provide the question, no explanation: {text}"

E Fine-tuning Model Training Details

Republican 1976-84 Train Validation

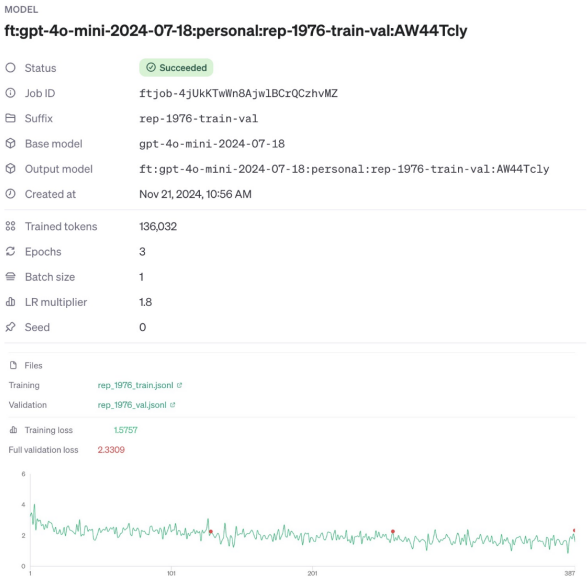

Figure 10: Training details for finetuning ChatGPT-4o mini Republican 1976-84

Democat 1976-84 Train Validation

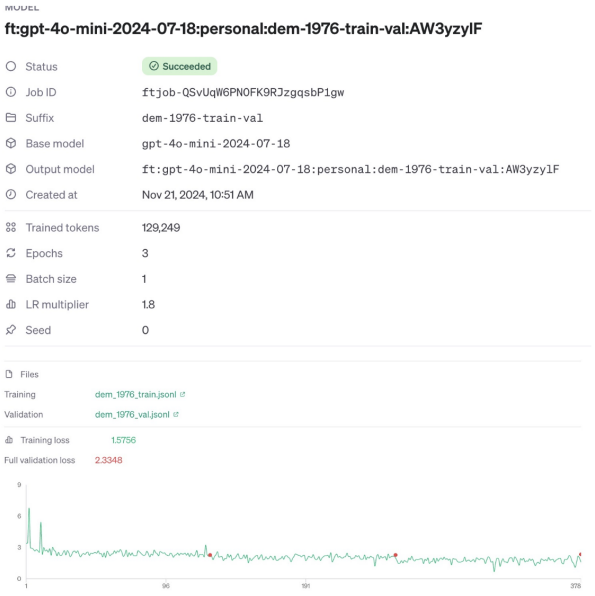

Figure 11: Training details for finetuning ChatGPT-4o mini Democrat 1976-84

Republican 2016-24 Train Validation

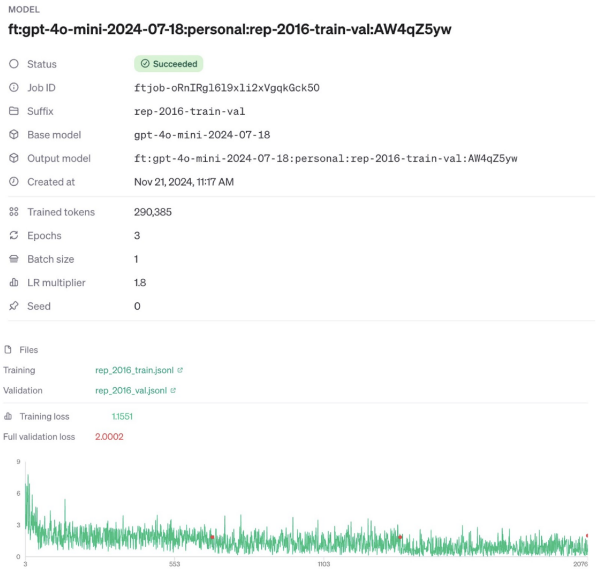

Figure 12: Training details for finetuning ChatGPT-4o mini Republican 2016-24

Democrat 2016-24 Train Validation

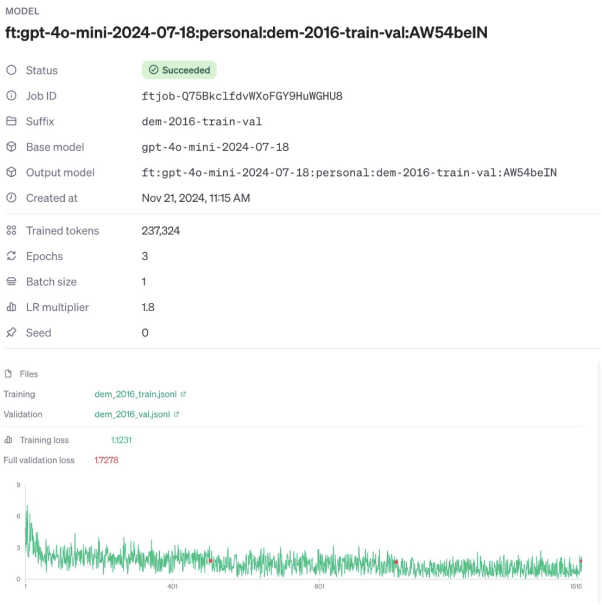

Figure 13: Training details for finetuning ChatGPT-4o mini Democrat 2016-24

F System Instructions for the Fine-Tuned LLMs, Representing Simulated Presidential Candidates

"You are a presidential candidate participating in a live presidential debate. Your role is to directly answer questions posed by the moderator and win over the American public."

G Questions Asked to Fine-Tuned Models

1. Why do you believe you are fit to be the next leader of the United States of America?
2. What is your overarching vision for the country?
3. What is your plan to manage inflation and ensure economic stability?
4. How will you create jobs and reduce unemployment across the nation?
5. What steps will you take to balance the federal budget and manage the national debt?
6. How will you ensure that all Americans have access to quality and affordable healthcare?
7. How will your administration address the pressing challenges of environmental sustainability?
8. What are your top priorities in regards to foreign policy?
9. How will you strengthen America's national defense and handle relations with adversarial nations?
10. What reforms would you propose to improve the education system?
11. How will you restore trust in government and ensure transparency in your administration?
12. What policies will you pursue to ensure equality and protect the rights of all Americans?
13. What is your plan to manage immigration in a way that balances national security, economic needs, and humanitarian responsibilities?
14. How will you ensure equal opportunities for marginalized communities in our country?

H Themes Top 10 Words for LDA Topics

Running the LDA model with 14 topics revealed the following overarching themes: Health Care Policy, National Policy Challenges, Civil Rights Issues, Economic Opportunities, Budget and Policy, Government Trust Issues, Environmental Sustainability, Education and Funding, Government and Taxes, Immigration and Jobs, Education and Students, Presidential Leadership, Policy and Strategy, National Leadership

Topic 1: health, people, one, care, would, program, think, thing, year
Topic 2: country, people, thats, president, used, take, way, law, inflation
Topic 3: right, american, people, opportunity, black, civil, minority, woman, year
Topic 4: going, job, inflation, make, country, people, million, new, sure
Topic 5: going, year, problem, country, america, budget, well, billion, president
Topic 6: government, health, care, trust, american, job, mr, well, restore
Topic 7: black, environmental, american, administration, sustainability, thats, time, seen, life
Topic 8: education, school, uhh, would, year, program, federal, budget, one
Topic 9: people, american, government, know, think, year, believe, state, country
Topic 10: people, america, country, job, many, come, think, back, say
Topic 11: bill, education, dont, people, let, school, year, college, need
Topic 12: president, job, people, one, year, would, thing, country, say
Topic 13: people, think, thats, well, wh, going, policy, make, one
Topic 14: people, president, ive, government, think, weve, thing, one, time