Linguistic Cues in Political Debates: Lexical Choices and Evaluative Framing in GPT-Generated Commentary

Anonymous ACL submission

Abstract

Political debates shape public opinion, but many readers increasingly encounter debate takeaways through AI-generated summaries and commentary rather than full broadcasts. This paper studies whether fine-grained lexical choices in debate transcripts can influence the evaluative framing of GPT-generated political commentary. We introduce controlled word-level interventions by replacing embedding-identified key lexical items with near-synonymous alternatives, allowing us to examine how minimally invasive lexical changes affect model-generated interpretation. Across U.S. presidential debate segments, we find that single-word substitutions can shift the stance of GPT-generated commentary, but these effects are probabilistic, context-dependent, and sensitive to the speaker's communicative goal. Substitutions are more likely to produce favorable stance shifts in critical contexts than in promotional ones, suggesting that GPT-generated commentary may amplify adversarial framing more readily than self-promotional framing. We further analyze which linguistic cue types are associated with substitution success, including sentiment direction, part of speech, and contextual suitability, and show that sentiment alone is insufficient to explain model responses. Finally, a small exploratory human evaluation suggests partial alignment between model-based stance shifts and reader preferences in promotional cases, while also revealing substantial subjectivity in human judgments. These findings highlight how candidates' word choices may be transformed when political discourse is filtered through GPT-generated commentary, and caution against treating model-internal stance metrics as direct substitutes for large-scale human audience evaluation.

1 Introduction

Public political debates play a central role in modern democracies, providing voters with opportunities to evaluate candidates' positions, competence, and credibility. Traditionally, research and campaign strategy have assumed a relatively direct communication pathway: candidates speak in debates, journalists and commentators interpret their performances, and human audiences form opinions based on these interpretations (Bennett, 2016; Entman, 2009). Within this framework, candidates' lexical and rhetorical choices function as *linguistic cues* that shape how messages are understood, evaluated, and remembered.

However, this communication pathway has changed substantially in recent years. An increasing portion of the public now encounters political debates indirectly, through LLM-generated summaries, commentary, or explanations rather than through full debate broadcasts or even human-written media coverage (Diakopoulos, 2019; Gilardi et al., 2023). Large Language Models are increasingly used to generate post-debate analyses, highlight key moments, summarize candidate positions, and answer user queries such as "Who won the debate?" or "What did the candidate say about healthcare?". As a result, many human readers are not reacting directly to candidates' original speech, but instead to LLM-mediated interpretations of that speech.

This shift introduces a new and largely unexamined layer of political communication: candidates' linguistic cues may first influence how an LLM interprets and comments on a debate, and only then reach human audiences through LLM-generated text. While prior work has extensively studied how rhetorical framing affects human audiences, much less is known about how subtle lexical choices affect machine-generated political commentary, and how these effects propagate to downstream readers who rely on such commentary.

Existing research on elections and debates has primarily focused on two areas. One line of work analyzes public reactions on social media follow-

ing debates or elections, such as sentiment dynamics, polarization, or incivility (Rossini et al., 2021; Xia et al., 2021). Another examines debate content itself, including linguistic structure, emotional expression, or gendered communication patterns (Fadilah and Kuswoyo, 2021; Boussalis et al., 2021). While these studies provide valuable insights into political discourse, they largely assume that audience perception is shaped directly by candidates' speech or by human-authored media interpretation. There is comparatively little work isolating how specific lexical choices in debate transcripts influence perception when interpretation is mediated by language models.

Studying this question directly with human commentators and audiences poses significant methodological challenges. It is impractical to ask journalists or large groups of voters to repeatedly evaluate many near-identical versions of the same debate speech that differ only by a single word. To address this limitation, we leverage Large Language Models as controlled experimental instruments. Recent work has shown that LLM outputs are highly sensitive to prompt wording and can exhibit systematic biases even when semantic content remains constant (Lin et al., 2025). This sensitivity suggests that LLMs respond to linguistic cues in ways that are comparable, at least in part, to human interpretation processes.

Biases in LLM behavior have been documented across multiple domains, including socioeconomic, political, and gender-related contexts (Kamruzzaman et al., 2024; Kotek et al., 2023; Röttger et al., 2024; Feng et al., 2023; Wambsganss et al., 2023). These biases emerge from statistical regularities in training data (Bender et al., 2021; Weidinger et al., 2022), and they shape how models associate particular words, frames, or tones with evaluative judgments. In the context of political debates, this means that an LLM may systematically favor or penalize certain lexical choices when generating commentary. By examining how small, controlled lexical changes affect LLM-generated evaluations, we can study how linguistic cues operate within this increasingly influential interpretive layer.

In this work, we conduct a series of controlled experiments to examine whether and how candidates' lexical choices in debate transcripts influence the stance of GPT-generated political commentary. We focus on fine-grained, word-level interventions, specifically, synonym substitutions at embedding-identified key positions, designed to minimally alter semantic content while modifying local linguistic cues. Using a two-stage LLM-mediated pipeline, we first generate professional-style commentary from debate transcripts and then score the stance of that commentary with respect to the speaker. We additionally conduct a small exploratory human evaluation to assess whether selected model-measured shifts correspond to perceptible differences in reader judgments.

Our analysis addresses three main questions. First, do subtle lexical substitutions in debate transcripts affect the tone and stance of GPT-generated political commentary, as reflected in the degree to which the commentary supports or criticizes the speaker's position? Second, which types of linguistic cues, such as sentiment, part of speech, contextual suitability, or word rarity, are most likely to be amplified or attenuated in GPT-generated evaluations? Third, do these effects vary depending on the communicative goal of the speaker, such as advancing one's own position versus criticizing an opponent? In an exploratory downstream analysis, we examine whether a subset of model-measured stance shifts aligns with human reader preferences over the generated commentaries.

By answering these questions, this study contributes to a growing understanding of political communication in LLM-mediated information environments. Rather than treating LLMs solely as passive tools or sources of bias, we examine their role as active intermediaries whose generated commentary can favor, neutralize, or undermine a speaker's stance based on subtle linguistic cues. At the same time, we distinguish between model-internal stance shifts and human audience effects: our main experiments measure changes in generated commentary, while our human evaluation provides limited exploratory evidence about whether such changes are noticeable to readers. Our findings therefore have implications for political actors, media platforms, and researchers seeking to understand how language choices may be transformed and transmitted through LLM-mediated channels.

2 Related Work

Prior work in political communication and discourse studies shows that subtle lexical choices can shape interpretation even when propositional content remains stable. In particular, framing research argues that describing the same event with different labels can shift salience and alter how

responsibility, legitimacy, or morality is inferred (Entman, 1993). More broadly, lexical choice, intensification, and rhetorical framing in political language can influence evaluative judgments, including perceptions of credibility, empathy, and moral positioning. These perspectives motivate our view that debate transcripts contain fine-grained linguistic cues with downstream evaluative consequences.

Within NLP and computational social science, a substantial body of work has analyzed political debates and elections at scale. Many studies focus on post-event public reactions on social media, including sentiment, polarization, and incivility (Rossini et al., 2021; Xia et al., 2021; Bossetta and Schmøkel, 2024). Other work examines debate transcripts directly, analyzing linguistic structure, rhetorical strategies, and speaker attributes such as gender-linked communication patterns (Fadilah and Kuswoyo, 2021; Boussalis et al., 2021). Together, these studies establish debate language as a useful signal for understanding political strategy and public discourse. Our focus, however, is not on public reaction or transcript analysis alone, but on the transformation process itself: how small lexical modifications in debate transcripts shift the evaluative stance of generated political commentary.

This question is especially important in light of recent work showing that LLM outputs can vary substantially under small prompt changes, even when semantic content remains similar (Lin et al., 2025). Such prompt sensitivity has been linked to instability, bias, and manipulability in generation systems, making controlled lexical edits a useful lens for studying model behavior. In parallel, a growing literature documents biases in LLMs across socioeconomic, gender, and political dimensions (Bender et al., 2021; Weidinger et al., 2022; Kotek et al., 2023; Röttger et al., 2024; Feng et al., 2023; Wambsganss et al., 2023). These findings suggest that models encode associations between linguistic cues and evaluative judgments, which may surface as systematic differences in generated stance. Our work brings these threads together in a political communication setting where debate transcripts serve as inputs and the outcome of interest is the evaluative tone of generated commentary rather than factual accuracy. This framing also connects to emerging datasets and tasks that pair transcripts with live or professional commentary (Chen et al., 2025), enabling more direct study of how political discourse is transformed into evaluation.

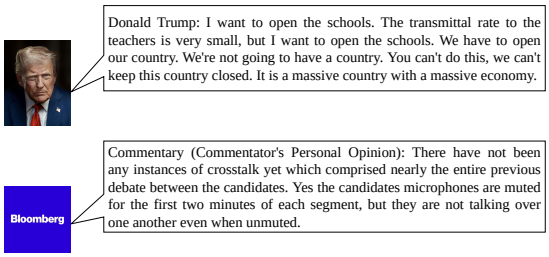

Figure 1: Examples of real-time expert commentary from Bloomberg in the dataset

3 Methodology

3.1 Task Definition and Overview

Our primary question is whether fine-grained linguistic cues in debate transcripts can shift the stance of GPT-generated political commentary in ways that benefit the speaker. This question is motivated by an increasingly common mediation pathway in which readers may consume debate takeaways through GPT-generated commentary rather than through full debate broadcasts.

The primary large-scale experiments in this paper are model-based. We operationalize the effect of linguistic cues by measuring how controlled lexical substitutions change the degree to which generated commentary supports or criticizes the speaker with respect to the speaker's communicative goal. We additionally conduct a small exploratory human evaluation in Section 4.3 to assess whether selected model-measured stance shifts correspond to perceptible differences in reader judgments.

For each transcript segment, we (i) identify key lexical positions, (ii) generate minimally edited variants through synonym substitution, (iii) generate commentary for each variant using GPT-4o, and (iv) score the stance of that commentary and compute success/failure relative to the speaker's goal. The resulting success labels should therefore be interpreted as model-based stance outcomes rather than direct measurements of population-level audience persuasion.

3.2 Dataset

We use the dataset introduced by Chen et al. (2025), which pairs transcripts with live professional commentary. It contains three subsets: U.S. presidential debates, FOMC press conferences, and corporate earnings calls. We focus on the U.S. presidential debate subset, which includes transcript segments,

professional commentaries, and commentary labels from ten debates between 2016 and 2024.

Each instance consists of a transcript segment (a contiguous speaking turn) and associated professional commentary. The professional commentary labels (supportive/critical/neutral) are used as reference signals to help prompt the LLM to produce commentary resembling professional political analysis. The original dataset comprises 2300 text segments across 11 unique labels. For this study, we specifically utilized the data annotated with the 'Commentator's Personal Opinion' label. On average, each case in this subset contains approximately 5 sentences and 60 words. Figure 1 shows an example of the dataset we used.

3.3 Identifying Key Lexical Positions

Our intervention is word-level synonym substitution. To keep changes minimally invasive while still meaningful, we only replace a small number of content words in each segment.

To select candidate replacement positions, we estimate word importance using an embedding-based masking approach with a pretrained Gemini Transformer model (Lee et al., 2025). Let q be a transcript segment and t a task prompt. We compute the embedding of the full segment as:

$$\mathbf{e}_{\text{full}} = f\Big(\text{mean_pool}(M(t \oplus q))\Big), \qquad (1)$$

where M is the encoder, $\oplus$ denotes concatenation, and $f(\cdot)$ is a linear projection.

For each candidate word w_i, we mask it in the segment to form q_{-i}, compute $\mathbf{e}_{-i}$, and define an importance score:

$$\Delta_i = \|\mathbf{e}_{\text{full}} - \mathbf{e}_{-i}\|. \qquad (2)$$

We select the top five words by Δ_i and restrict candidates to content words (nouns, verbs, adjectives, adverbs), excluding function words to avoid uninformative or ungrammatical replacements.

3.4 Synonym Substitution and Fluency Control

For each selected word w_i, we retrieve near-synonyms from Datamuse and construct transcript variants by replacing w_i with a candidate synonym s_{ij}. Each variant differs from the original transcript by exactly one lexical substitution. Because synonyms may be contextually incompatible or grammatically infelicitous, we post-process each variant with GPT-4o to correct grammar while preserving meaning. This yields a set of fluent transcript variants with minimal semantic deviation but potentially different lexical cueing.

3.5 Generating LLM Commentary

For each transcript variant (including the original), we prompt GPT-4o to produce a short piece of evaluative political commentary in the style of professional debate analysis. We condition the model with a fixed instruction template and provide examples from the dataset as style anchors. The output is a commentary text that may implicitly favor the speaker, criticize the speaker, or remain neutral.

3.6 Commentary Stance Scoring

To evaluate how a lexical substitution changes the generated stance, we score each commentary along a three-way stance scale with respect to the speaker:

$$y \in \{\textsc{Supportive}, \textsc{Neutral}, \textsc{Critical}\}.$$

We obtain this stance score by prompting GPT-4o (in a separate scoring prompt) to classify the generated commentary as supportive/neutral/critical toward the speaker and to output calibrated probabilities:

$$p = (p_{\text{sup}}, p_{\text{neu}}, p_{\text{crit}}), \quad p_{\text{sup}} + p_{\text{neu}} + p_{\text{crit}} = 1.$$

We then define a scalar Commentary Stance Score:

$$S = p_{\text{sup}} - p_{\text{crit}}, \qquad (3)$$

where higher S indicates more supportive commentary and lower S indicates more critical commentary. This score is computed for the original transcript and each substituted variant.

3.7 Goal Annotation

Debate segments often serve different rhetorical purposes. We therefore label each transcript segment with a communicative goal:

$$g \in \{\textsc{Promotional}, \textsc{Critical}\}.$$

We first assign sentence-level goals (promotional vs. critical) using an instruction prompt to GPT-4o, then aggregate to the segment level by majority vote over sentences. Segments that contain both goals are assigned the dominant goal.

3.8 Success, Failure, and Stability

We define success at the level of a single substitution by comparing the stance score of the modified transcript S_{mod} to that of the original transcript S_{orig}. Let

$$\Delta S = S_{\text{mod}} - S_{\text{orig}}. \quad (4)$$

A substitution is classified as:

- **Success** if $\Delta S > \tau$,
- **Failure** if $\Delta S < -\tau$,
- **Stable** otherwise,

where τ is a small stability threshold (we use $\tau = 0$ in our analysis).

This definition is applied uniformly to both PROMOTIONAL and CRITICAL segments. In promotional contexts, higher S indicates stronger support for the speaker's self-presentation. In critical contexts, higher S reflects commentary that is more supportive of the speaker's critique of an opponent (i.e., frames the attack as more compelling or less dismissible).

The overall success rate is the proportion of substitutions labeled **Success** among all tested substitutions, and serves as the primary metric in our Replacement Success Measurement analysis.

3.9 Linguistic Analyses of Substitutions

To interpret why some substitutions succeed or fail, we characterize each replacement along multiple linguistic dimensions. We compute lexicon-based sentiment using SentiWordNet (Esuli and Sebastiani, 2006) (Pos/Neg/Obj) for both the original word and the substitute, and we additionally obtain LLM-based connotation judgments via a fixed scoring prompt. We further annotate each substitution with part-of-speech, contextual suitability (rated by GPT-4o on a four-level scale), and word rarity using Wordfreq. We then analyze how these attributes correlate with ΔS and with success/failure outcomes.

4 Results and Analysis

4.1 Overall Replacement Outcomes

We begin by reporting the overall effect of lexical substitutions under the goal-aligned success definition introduced in Section 3. Recall that a substitution is labeled **Success** if it shifts the commentary stance score in the desired direction ($\Delta S > \tau$), **Failure** if it shifts in the opposite direction ($\Delta S < -\tau$), and **Stable** otherwise.

Table 1 summarizes the distribution of success, stability, and failure across all tested substitutions, as well as separately for promotional and critical segments. Across all debate segments, 40.9% of substitutions are classified as successful. This indicates that even a single-word change can alter the stance of GPT-generated political commentary in a direction favorable to the speaker. However, the effect is not uniformly beneficial: 46.4% of substitutions result in failure, meaning that the same type of lexical intervention can also lead to commentary that is less supportive of the speaker's intended framing. The remaining 12.8% of substitutions are classified as stable, producing no directional change under our threshold.

These results suggest that word choice functions as a probabilistic rather than deterministic lever in LLM-mediated political interpretation. Lexical substitutions can matter, but their effects are contingent on the local context and can easily backfire when the substituted word disrupts the speaker's rhetorical framing.

A notable asymmetry emerges when separating substitutions by communicative goal. For promotional segments, only 27.2% of substitutions are successful, while 60.3% result in failure. In contrast, substitutions in critical segments achieve a substantially higher success rate of 58.3%, with a lower failure rate of 28.6%. This gap suggests that lexical choices may be more readily amplified by the commentary model when they sharpen or reinforce criticism of an opponent than when they attempt to improve self-presentation.

One possible explanation is that critical discourse provides a clearer evaluative target: forceful wording can directly signal disapproval, which the LLM may convert into commentary that treats the speaker's attack as more compelling. Promotional discourse, by contrast, requires a more delicate balance between positivity, credibility, and restraint. Overly positive or stylistically mismatched substitutions may be interpreted as exaggeration or lack of substance, leading generated commentary to respond skeptically.

Taken together, these results establish two descriptive patterns. First, single-word substitutions are a non-trivial source of variation in GPT-generated political commentary. Second, the direction of this variation is highly context-dependent: lexical editing is not a reliable optimization strat-

Goal	Success	Stable	Failure	Total
Overall	2,600 (40.86%)	813 (12.78%)	2,950 (46.36%)	6,363
Promotional	972 (27.23%)	446 (12.49%)	2,152 (60.28%)	3,570
Critical	1,628 (58.29%)	367 (13.14%)	798 (28.57%)	2,793

Table 1: Overall replacement outcomes under the goal-aligned success definition. A substitution is labeled **Success** if $\Delta S > \tau$, **Failure** if $\Delta S < -\tau$, and **Stable** otherwise.

Sentiment	Shift	Success Rate (SentiWordNet)	Success Rate (LLM)
Pos.	↑	33.80%	40.38%
	↓	50.00%	43.41%
Neg.	↑	39.13%	42.98%
	↓	40.54%	41.69%

Table 2: Success rates by sentiment-shift direction under two sentiment estimators. ↑ and ↓ indicate an increase or decrease in the corresponding sentiment score of the substituted word relative to the original word (e.g., Pos.↑ = more positive).

Sentiment	Shift	Noun	Verb	Adv.
Pos.	↑	27.27%	37.82%	36.76%
	↓	36.89%	47.15%	36.84%
Neg.	↑	43.69%	41.91%	48.57%
	↓	48.20%	37.92%	44.64%

Table 3: Success rates by sentiment-shift direction under the LLM-based sentiment estimator, stratified by part of speech.

egy by itself, but a sensitive intervention whose effect depends on the speaker's goal and the rhetorical environment. In the following subsections, we analyze how specific linguistic properties of substitutions, such as sentiment, part of speech, and contextual suitability, help explain when substitutions succeed or fail.

4.2 Which Lexical Factors Correlate with Success?

Table 1 shows that lexical substitutions produce a substantial success rate overall, but also a high failure rate, and that the success rate differs sharply between promotional and critical segments. We next ask which lexical properties are associated with successful substitutions. Concretely, we analyze how ΔS and success/failure outcomes vary with (i) the sentiment shift of the substituted word, (ii) part of speech, and (iii) contextual suitability.

4.2.1 Sentiment Shift and Success Rate

We first examine whether the direction of lexical sentiment change alone is predictive of substitution success. Intuitively, replacing a word with a more positive or less negative synonym might be expected to increase the likelihood that generated commentary favors the speaker. To test this assumption, we analyze success rates as a function of sentiment shift under two different sentiment estimators.

We measure word-level sentiment using both a lexicon-based and a model-based approach. First, we use *SentiWordNet*, a widely used sentiment lexicon built on WordNet 3.0 (Esuli and Sebastiani, 2006). Each word is assigned three real-valued scores: positive (Pos), negative (Neg), and objective (Obj), each in $[0, 1]$ and summing to 1. Second, we obtain an LLM-based sentiment estimate by prompting GPT-4o to assess the connotation of a word in isolation, returning probabilities over the same three categories. This second measure is intended to capture sentiment associations that may not be well represented in static lexicons but are salient to the LLM's own interpretive behavior.

For each substitution, we compare the sentiment scores of the original word and its synonym and categorize the change as an increase or decrease along each sentiment dimension. Table 2 reports success rates for positive and negative sentiment shifts under both sentiment estimators.

Several patterns emerge. First, sentiment shift alone is not a strong or monotonic predictor of success. Across both estimators, increasing positivity does not consistently yield higher success rates than decreasing positivity. Under SentiWordNet, substitutions that reduce positivity achieve a higher success rate (50.0%) than those that increase positivity (33.8%). Under the LLM-based estimator, the gap is smaller but points in the same direction. This suggests that more positive wording is not automatically rewarded in GPT-generated political commentary; in some contexts, reduced positivity may make the statement appear more restrained or credible.

Second, negative sentiment shifts show relatively symmetric behavior. Making a word more negative or less negative yields comparable success rates under both estimators. This indicates that the presence of negativity alone does not determine whether a substitution succeeds. Rather, negative wording appears to help only when it is compatible

Sentiment	Shift	Highly Appropriate	Appropriate	Inappropriate	Highly Inappropriate
Pos.	↑	30.00%	42.59%	42.40%	30.00%
	↓	42.86%	54.55%	42.14%	37.33%
Neg.	↑	50.00%	36.07%	48.08%	41.73%
	↓	50.00%	41.96%	40.00%	42.81%

Table 4: Success rates by sentiment-shift direction (LLM-based estimator) stratified by contextual suitability.

with the broader rhetorical goal and local context.

Finally, while the absolute values differ slightly between SentiWordNet and the LLM-based estimator, both measures support the same general conclusion: lexical sentiment direction by itself is insufficient to explain substitution success. Sentiment matters, but its effect is mediated by other lexical properties, including part of speech and contextual suitability. We therefore turn next to analyses that incorporate these additional factors.

4.2.2 Part of Speech Effects

The sentiment analysis above treats substitutions as if lexical sentiment shifts have uniform effects regardless of where they occur in a sentence. We next test whether the impact of a sentiment shift depends on part of speech (POS), since nouns, verbs, and adverbs play different rhetorical roles in debate language. Nouns often anchor topics and entities, verbs encode agency and action, and adverbs frequently modulate intensity and speaker attitude. These functional differences may lead GPT-generated commentary to react differently to the same type of sentiment shift depending on POS.

Table 3 reports success rates under the LLM-based sentiment estimator, stratified by POS. Two patterns stand out. First, verbs show the strongest sensitivity to positivity attenuation: when the substituted word becomes less positive (Pos.↓), verb substitutions achieve the highest success rate among the positive-shift cases (47.15%), substantially higher than noun substitutions under the same shift (36.89%). Combined with the weaker performance of Pos.↑ for nouns (27.27%), this suggests that increasing positivity on content nouns is particularly risky, whereas moderating positivity on verbs may improve perceived credibility or reduce exaggeration in generated commentary. Second, negative shifts interact with POS in a goal-consistent way. For adverbs, making a word more negative (Neg.↑) yields the highest success rate in the table (48.57%). This is consistent with the role of adverbs as intensity markers: stronger negative adverbs can sharpen criticism or signal conviction, which the commentary model may interpret as rhetorically effective. In contrast, verb substitutions exhibit a lower success rate when Neg.↓ (37.92%), indicating that weakening negativity in action framing may reduce the perceived forcefulness of a critique. Overall, these POS-stratified results reinforce the conclusion from Table 2 that sentiment shifts alone do not fully explain success. Instead, the effect of sentiment depends on how the modified word functions in the sentence. This motivates our next analyses on contextual suitability, which further characterize when lexical substitutions are amplified or attenuated by GPT-generated commentary.

4.2.3 Contextual Suitability

The analyses above suggest that sentiment shift alone does not deterministically explain substitution success. We next examine contextual suitability, which captures whether a synonym is pragmatically and stylistically appropriate in its local discourse context. Even near-synonyms can differ in register, collocations, and implied presuppositions; such mismatches may cause the generated commentary to react negatively regardless of sentiment direction. For each substitution, we obtain a contextual suitability label from GPT-4o by prompting it to judge how appropriate the substituted word is in the original sentence context. We use a four-level scale: HIGHLY APPROPRIATE, APPROPRIATE, INAPPROPRIATE, and HIGHLY INAPPROPRIATE. We then compute success rates within each suitability level and sentiment-shift category.

Table 4 reports success rates stratified by suitability and sentiment-shift direction (LLM-based sentiment estimator). Two trends are notable. First, higher contextual suitability is generally associated with higher success rates, consistent with the intuition that locally well-formed substitutions are more likely to be interpreted favorably in generated commentary. Second, the interaction between sentiment shift and suitability is asymmetric: in multiple cells, decreasing positivity (Pos.↓) is more successful than increasing positivity (Pos.↑), especially when the substitute word is judged appropriate. This pattern aligns with earlier observa-

tions that overly positive wording can be penalized, while restrained phrasing that preserves contextual fit is more likely to succeed. For negative sentiment, both Neg.↑ and Neg.↓ can achieve high success when the substitution is contextually appropriate, indicating that negativity can be effective when it matches local discourse expectations (e.g., criticism segments or emphatic modifiers). In contrast, when substitutions are highly inappropriate, success rates drop and become less interpretable, suggesting that contextual mismatch dominates sentiment direction. These results reinforce that contextual suitability is a primary constraint on successful lexical editing: sentiment shifts appear to help mainly when the substitute is already compatible with the local context.

4.3 Exploratory Human Evaluation: Audience Preferences over GPT-Generated Commentary

To complement the model-based stance analysis, we conduct a small-scale exploratory human evaluation. The purpose of this evaluation is not to establish population-level persuasion effects, but to examine whether stance shifts measured by our model-based metric correspond to perceptible differences in reader judgments of GPT-generated commentary. We sampled 200 paired cases, evenly balanced across four conditions: PROMOTIONAL SUCCESS, PROMOTIONAL FAILURE, CRITICAL SUCCESS, and CRITICAL FAILURE. For each case, annotators compared the pre-substitution and post-substitution commentaries and selected which version left a more favorable overall impression of the speaker, or whether no clear preference was present. Annotators evaluated the generated commentaries rather than the original transcripts, allowing us to focus on the downstream interpretive layer that readers may encounter in LLM-mediated political information settings.

In promotional contexts, we observe the clearest alignment between the model-based success label and human preference. For cases classified as PROMOTIONAL SUCCESS, annotators preferred the post-substitution commentary in 65% of judgments. By contrast, this rate decreased to 46% for PROMOTIONAL FAILURE cases. This contrast suggests that model-measured improvements in promotional stance can sometimes correspond to more favorable human impressions, although the alignment is far from deterministic. In critical contexts, the pattern is less clear. Post-substitution commentaries were preferred in 48% of CRITICAL SUCCESS cases and 49% of CRITICAL FAILURE cases. These near-comparable rates indicate that the model-based success distinction does not translate into a consistent aggregate human preference in critical rhetoric. One possible explanation is that readers differ in how they evaluate aggressive or adversarial framing: stronger criticism may appear persuasive to some readers but excessive or unfair to others. Inter-annotator agreement was modest, with Cohen's $\kappa = 0.24$. We therefore interpret these results cautiously. The relatively low agreement indicates that judgments of political commentary favorability are subjective and may depend on individual norms regarding tone, fairness, and rhetorical force. Rather than treating the human evaluation as conclusive evidence of audience persuasion, we view it as an exploratory check on whether model-measured stance shifts can be noticeable to human readers.

5 Conclusion

This paper studies whether fine-grained lexical choices in presidential debate transcripts can affect the evaluative framing of GPT-generated political commentary. Through controlled single-word synonym substitutions, we show that lexical edits can shift model-generated stance, but that these effects are probabilistic and highly context-dependent. In particular, substitutions are more successful in critical contexts than in promotional ones, suggesting that GPT-generated commentary may respond more readily to adversarial framing than to self-promotional framing. Our analyses further show that sentiment direction alone does not explain substitution success. The effect of a lexical change depends on how the substituted word functions in context, including its part of speech and contextual suitability. A small exploratory human evaluation provides limited downstream support, with the clearest alignment appearing in PROMOTIONAL SUCCESS cases, while also revealing substantial subjectivity in reader judgments. Overall, these findings suggest that subtle lexical cues can be amplified, weakened, or transformed when political discourse is filtered through GPT-generated commentary. At the same time, the high failure rate, modest human agreement, and model-specific design caution against interpreting model-internal stance shifts as direct evidence of population-level audience persuasion.

Limitations

Our study has several limitations and points toward future work. First, the primary evaluation focuses on LLM-generated commentary stance rather than direct human audience attitudes. Although we include a small exploratory human evaluation, it involved only two annotators and a limited sample of 200 paired cases. The modest inter-annotator agreement further indicates that perceived favorability of political commentary is subjective and reader-dependent. Future work should recruit larger and more politically diverse participant samples, report fuller preference distributions, and examine whether observed commentary differences translate into stable changes in audience attitudes.

Second, our interventions are restricted to single-word synonym substitutions. This design allows for controlled comparisons, but richer edits, such as multi-word paraphrases, discourse-level reframing, or changes in argument structure, may produce stronger effects while also introducing more semantic confounds. Future studies should examine how different levels of linguistic intervention affect LLM-generated political interpretation.

Third, our main pipeline relies primarily on GPT-4o for commentary generation, stance scoring, and several linguistic annotations. This design provides consistency across experimental stages, but it may also concentrate model-specific biases. Our preliminary cross-model analysis suggests that substitution outcomes vary across LLMs, but the analysis is limited in scale and should not be treated as a full robustness evaluation. Future work should replicate the experiment across additional commercial and open-source models, prompting strategies, languages, and political domains.

Finally, our success metric captures directional change in model-generated stance, not normative quality, factuality, or democratic desirability. A substitution that makes commentary more favorable to a speaker is not necessarily more accurate, fair, or informative. Future research should therefore jointly evaluate stance shifts, factual consistency, rhetorical fairness, and reader trust when studying LLM-mediated political communication.

References

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

W Lance Bennett. 2016. *News: The politics of illusion*. University of Chicago Press.

Michael Bossetta and Rasmus Schmøkel. 2024. Cross-platform emotions and audience engagement in social media political campaigning: comparing candidates' facebook and instagram images in the 2020 us election. In *Dissonant Public Spheres*, pages 35–55. Routledge.

Constantine Boussalis, Travis G Coan, Mirya R Holman, and Stefan Müller. 2021. Gender, candidate emotional expression, and voter reactions during televised debates. *American Political Science Review*, 115(4):1242–1257.

Chung-Chi Chen, Huan-Wen Ho, Yu-Yu Chang, Ming-Hung Wang, Ramon Ruiz-Dolz, Chris Reed, Ichiro Kobayashi, Yusuke Miyao, and Hiroya Takamura. 2025. Live commentary planning and generation. In *Proceedings of the 18th International Natural Language Generation Conference: Generation Challenges*, pages 37–43, Hanoi, Vietnam. Association for Computational Linguistics.

Nicholas Diakopoulos. 2019. *Automating the news: How algorithms are rewriting the media*. Harvard University Press.

Robert M. Entman. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4):51–58.

Robert M Entman. 2009. *Projections of power: Framing news, public opinion, and US foreign policy*. University of Chicago Press.

Andrea Esuli and Fabrizio Sebastiani. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. In *LREC*, volume 6, pages 417–422.

Rani Fadilah and Heri Kuswoyo. 2021. Transitivity analysis of presidential debate between trump and biden in 2020. *Linguist. Lit. J*, 2(2):98–107.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair nlp models. *arXiv preprint arXiv:2305.08283*.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Mahammed Kamruzzaman, Hieu Minh Nguyen, and Gene Louis Kim. 2024. " global is good, local is bad?": Understanding brand bias in llms. *arXiv preprint arXiv:2406.13997*.

Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of the ACM collective intelligence conference*, pages 12–24.

Jinhyuk Lee, Feiyang Chen, Sahil Dua, Daniel Cer, Madhuri Shanbhogue, Iftekhar Naim, Gustavo Hernández Ábrego, Zhe Li, Kaifeng Chen, Henrique Schechter Vera, Xiaoqi Ren, Shanfeng Zhang, Daniel Salz, Michael Boratko, Jay Han, Blair Chen, Shuo Huang, Vikram Rao, Paul Suganthan, and 28 others. 2025. Gemini embedding: Generalizable embeddings from gemini.

Weiran Lin, Anna Gerchanovsky, Omer Akgul, Lujo Bauer, Matt Fredrikson, and Zifan Wang. 2025. Llm whisperer: An inconspicuous attack to bias llm responses. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–24.

Patricia Rossini, Jennifer Stromer-Galley, and Feifei Zhang. 2021. Exploring the relationship between campaign discourse on facebook and the public's comments: A case study of incivility during the 2016 us presidential election. *Political Studies*, 69(1):89–107.

Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Kirk, Hinrich Schütze, and Dirk Hovy. 2024. Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15295–15311.

Thiemo Wambsganss, Xiaotian Su, Vinitra Swamy, Seyed Neshaei, Roman Rietsche, and Tanja Käser. 2023. Unraveling downstream gender bias from large language models: A study on ai educational writing assistance. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10275–10288.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, and 4 others. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 214–229, New York, NY, USA. Association for Computing Machinery.

Ethan Xia, Han Yue, and Hongfu Liu. 2021. Tweet sentiment analysis of the 2020 us presidential election. In *Companion proceedings of the web conference 2021*, pages 367–371.

A Appendix

A.1 Human Evaluation Process

We recruited two graduate students to serve as human evaluators for the LLM-generated commentary. Each evaluation item consisted of a pair of generated commentaries corresponding to the pre-substitution and post-substitution transcript variants. The original transcripts were not shown to annotators, so that the task focused on the downstream commentary that readers would encounter in an LLM-mediated information setting.

Sentiment	Shift	Success Rate (SentiWordNet)	Success Rate (LLM)
Pos.	↑	11.55%	12.98%
	↓	23.68%	20.63%
Neg.	↑	15.21%	12.31%
	↓	10.81%	11.14%

Table 5: Success rates with $\tau = 0.05$ by sentiment-shift direction under two sentiment estimators. ↑ and ↓ indicate an increase or decrease in the corresponding sentiment score of the substituted word relative to the original word (e.g., Pos.↑ = more positive).

Sentiment	Shift	Noun	Verb	Adv.
Pos.	↑	27.27%	14.28%	04.69%
	↓	15.15%	21.29%	15.62%
Neg.	↑	16.77%	11.53%	15.69%
	↓	09.41%	10.48%	09.17%

Table 6: Success rates with $\tau = 0.05$ by sentiment-shift direction under the LLM-based sentiment estimator, stratified by part of speech.

The evaluators were asked to perform a comparative favorability assessment. Specifically, for each pair, they selected which commentary left a more favorable overall impression of the speaker, or selected that no clear preference was present. Participants were informed that the commentaries were AI-generated and were based on U.S. presidential debate content. They were instructed to read the commentaries as general audience members encountering political news or post-debate analysis, while also paying attention to subtle differences in tone, confidence, and evaluative framing.

The sampled evaluation set contained 200 paired cases, evenly balanced across PROMOTIONAL SUCCESS, PROMOTIONAL FAILURE, CRITICAL SUCCESS, and CRITICAL FAILURE conditions. These conditions were determined using the model-based stance metric described in Section 3. The human evaluation was therefore designed as an exploratory downstream check on model-measured stance shifts, rather than as a population-level study of persuasion. The detailed results are reported in Section 4.3.

Sentiment	Shift	Highly Appropriate	Appropriate	Inappropriate	Highly Inappropriate
Pos.	↑	10.00%	18.18%	08.33%	28.57%
	↓	30.00%	30.77%	21.46%	15.58%
Neg.	↑	0.00%	08.96%	16.28%	12.08%
	↓	40.00%	12.30%	08.06%	11.93%

Table 7: Success rates with $\tau = 0.05$ by sentiment-shift direction (LLM-based estimator) stratified by contextual suitability.

A.2 Success Rate with Stability Threshold

As noted in 3.8, the stability threshold τ was initially set to 0. While our original assumption was that any shift in stance would influence audience perception, we acknowledge that minor fluctuations may fall within the margin of error rather than representing a meaningful shift in opinion. Therefore, in this analysis, we set $\tau = 0.05$, stipulating that the difference in success rates before and after substitution must exceed 0.05 to be categorized as a successful modification.

The results in Tables 5, 6, and 7 show that applying a nonzero stability threshold substantially reduces the number of substitutions classified as successful. This indicates that many substitutions labeled as successes under $\tau = 0$ correspond to relatively small stance-score changes. The main results should therefore be interpreted as evidence of model sensitivity to lexical substitutions, rather than as evidence that every successful substitution produces a large or practically meaningful shift.

This thresholded analysis reinforces the importance of the stability parameter τ. A stricter threshold distinguishes larger stance shifts from marginal score fluctuations, but it also reduces the number of cases classified as successful. Future work should conduct a more systematic sensitivity analysis over multiple threshold values and examine whether larger model-measured stance shifts are more likely to correspond to consistent human preferences.

A.3 Significance Tests

We conducted categorical significance tests to examine whether lexical substitution is associated with changes in model-assigned commentary stance. Specifically, we compare the stance labels assigned to commentaries generated from the original transcripts and from the substituted transcripts. As shown in Table 8, both McNemar's test and the Stuart-Maxwell test reject the null hypothesis of no marginal change.

These tests indicate that lexical substitutions are associated with statistically detectable changes in model-assigned commentary stance. Importantly, they should not be interpreted as significance tests of human audience perception. Rather, they provide additional evidence that the generated commentary pipeline is sensitive to controlled lexical changes in the input transcript.

	Statistic	P-value
McNemar's test	296.0430	7.14e-64
Stuart-Maxwell test	284.4141	8.20e-64

Table 8: Significance tests comparing model-assigned commentary stance before and after lexical substitution.

	GPT	Gemini	Claude
Success	50%	35.00%	45.00%
Fail	50%	59.00%	52.00%
Stable	0%	6.00%	3.00%

Table 9: Preliminary cross-model check on a GPT-balanced subset. The GPT column reflects the sampling design rather than an independent comparison.

A.4 Preliminary Cross-Model Check

To assess whether lexical sensitivity is specific to GPT-4o commentary generation, we conduct a small preliminary cross-model check with two additional commercial LLMs, Gemini and Claude. We sample 200 instances from the primary GPT-4o experiment, consisting of 100 cases labeled as successful and 100 cases labeled as failed under the GPT-4o-based pipeline.

For each instance, Gemini and Claude are prompted to generate commentaries for the original and substituted transcript variants using the same commentary-generation setup. We then apply the same GPT-based stance-scoring procedure to evaluate whether the post-substitution commentary shifts in the goal-aligned direction. Because the subset is intentionally balanced according to the GPT-4o pipeline, the GPT column in Table 9 reflects the sampling construction rather than an independent model comparison.

Table 9 suggests that substitution outcomes vary across commentary models. Gemini produces a lower success rate and a higher failure rate than the GPT-balanced subset, while Claude falls closer to the balanced distribution but still differs from

GPT. Both Gemini and Claude also produce a small number of stable cases, indicating that some substitutions that shift GPT-generated commentary do not produce a directional stance change when other models generate the commentary.

These results should be interpreted cautiously. The analysis is based on a small GPT-selected subset, and the stance-scoring procedure remains GPT-based. Therefore, this experiment does not provide a full robustness evaluation across models. Rather, it offers preliminary evidence that lexical substitution effects may depend on the commentary-generation model, motivating future work with larger samples, model-specific scoring, and cross-model evaluation designs.