Linguistic Cues in Political Debates: Lexical Choices and Ideological Framing in LLM-Generated Commentary

Anonymous ACL submission

Abstract

Political debates shape public opinion, but many viewers now encounter debates through LLM-generated summaries and commentary rather than full broadcasts. We study whether fine-grained lexical and rhetorical choices in debate transcripts can systematically influence an LLM's evaluative commentary, and whether such shifts propagate to downstream audience impressions. We introduce controlled word-level interventions by replacing embedding-identified key lexical items with near-synonymous alternatives. Results show that even minimally invasive substitutions can reliably shift the tone of LLM-generated commentary and alter perceived audience sentiment. We further analyze which linguistic cue types are most likely to be amplified or attenuated in LLM-mediated political discourse, highlighting how candidates' word choices can indirectly shape public perception when LLM commentary serves as a primary interpretive layer.

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. 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. 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 following 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 audience perception through LLM-generated commentary. We focus on fine-grained, word-level interventions—specifically, synonym substitutions at embedding-identified key positions—designed to minimally alter semantics while modifying linguistic cues. Using a two-stage LLM-mediated pipeline, we first generate professional-style commentary from debate transcripts and then simulate audience impressions based on that commentary. This setup allows us to trace how linguistic cues propagate through the sequence *transcript* $\rightarrow$ *LLM commentary* $\rightarrow$ *audience impression*.

Our analysis addresses three main questions. First, do subtle lexical substitutions in debate transcripts systematically affect the tone and stance of LLM-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 LLM-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?

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 systematically favor, neutralize, or undermine a speaker's stance based on subtle linguistic cues. 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

2.1 Linguistic Cues, Framing, and Evaluative Language in Politics

A long line of work in political communication and discourse studies has shown that subtle lexical choices can function as cues that guide interpretation, often without changing the underlying propositional content. Framing research argues that describing the same event with different labels can shift which considerations become salient and how responsibility, legitimacy, or morality is inferred (Entman, 1993). More broadly, prior work suggests that lexical choice, intensification, and rhetorical framing in political text can influence evaluative judgments, shaping how speakers are perceived in terms of credibility, empathy, and moral positioning, even when the underlying message remains stable. These perspectives motivate our view that debate transcripts contain fine-grained linguistic cues that can influence downstream evaluations. Our work extends this line of inquiry to a setting where interpretation is increasingly mediated by LLM-generated political commentary. In such settings, the relevant question is not only how humans react to candidates' language, but also how models translate linguistic cues into evaluative judgments and how those judgments may be consumed as secondary political information.

2.2 Computational Studies of Debates and Political Discourse

Within NLP and computational social science, a substantial body of work studies political debates and elections through large-scale text analysis. 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). Another set of studies analyzes debate transcripts directly, examining linguistic structure, rhetorical strategies, and speaker attributes such as gender-linked patterns (Fadilah and Kuswoyo, 2021; Boussalis et al., 2021). These works establish that debate language is a useful signal for analyzing political strategy and public discourse. Our focus differs in previous studies we study the transformation itself: how small lexical modifications in debate transcripts shift the stance of generated political commentary, operationalized through LLM outputs.

2.3 LLMs, Prompt Sensitivity, and Political Bias in Generated Text

Recent work has documented that LLM outputs can vary substantially under small prompt changes, even when semantics are similar (Lin et al., 2025). This prompt sensitivity has been studied as a source of instability, bias, and manipulability in generation systems, and it motivates using controlled lexical edits as a lens for analyzing model behavior. In parallel, a growing literature has examined 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 can surface as systematic differences in generated stances. Our work leverages these properties in a targeted political communication setting: debate transcripts serve as prompts, and the outcome of interest is not factual correctness but the evaluative stance expressed in generated commentary. 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 discourse is transformed into evaluation.

3 Methodology

3.1 Task Definition and Overview

Our primary question is whether fine-grained linguistic cues in debate transcripts can systematically shift the stance of LLM-generated political commentary in ways that benefit the speaker. This is motivated by an increasingly common mediation pathway in which human readers consume debate takeaways through LLM-generated commentary rather than through full debate broadcasts. In our experiments, we do not recruit human participants. Instead, 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.

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.

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.

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:

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

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:

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

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:

$$y \in \{\text{SUPPORTIVE}, \text{NEUTRAL}, \text{CRITICAL}\}.$$

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:

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

We then define a scalar Commentary Stance Score:

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

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:

$$g \in \{\text{PROMOTIONAL}, \text{CRITICAL}\}.$$

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 the stance score of the original transcript S_{orig}.

For a PROMOTIONAL segment, the desired direction is to increase support for the speaker:

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

A substitution is:

- **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).

For a CRITICAL segment, the speaker aims to undermine the opponent. Since our commentary is generated about the speaker's utterance, we operationalize success as producing commentary that frames the speaker's attack as more compelling, which typically manifests as commentary that is more supportive of the speaker's critique (or less dismissive of it). Concretely, we use the same stance score relative to the speaker and define:

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

This aligns the success definition with our empirical objective: whether a lexical cue increases the extent to which model-generated commentary favors the speaker's intended framing.

The overall success rate is the proportion of substitutions labeled **Success** among all tested substitutions. This success rate is the metric reported in our Replacement Success Measurement analysis.

Goal	Success	Stable	Failure	Total
Overall	2,600 (40.86%)	813 (12.78%)	2,950 (46.36%)	6,363
Promotional	972 (31.91%)	446 (14.64%)	1,628 (53.45%)	3,046
Critical	1,628 (49.08%)	367 (11.06%)	1,322 (39.86%)	3,317

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.

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 effectiveness 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, approximately 40.9% of substitutions are successful, indicating that even a single-word change has a substantial probability of shifting LLM-generated political commentary in a direction favorable to the speaker. This result supports our central hypothesis that fine-grained linguistic cues can meaningfully influence model-mediated political interpretation.

At the same time, nearly half of the substitutions (46.4%) result in failure, highlighting that lexical manipulation is inherently risky: poorly chosen substitutions can backfire and lead to commentary that is less supportive of the speaker's intended framing. The remaining 12.8% of substitutions are classified as stable, producing negligible change in commentary stance. Taken together, these results suggest that word choice offers a probabilistic rather than deterministic lever for influencing LLM-generated commentary.

A notable asymmetry emerges when separating substitutions by communicative goal. For promotional segments, only 31.9% of substitutions are successful, while more than half (53.5%) result in failure. In contrast, substitutions in critical segments achieve a substantially higher success rate of 49.1%, with a lower failure rate of 39.9%. This gap suggests that lexical choices are more effective when used to sharpen or reinforce criticism of an opponent than when used to enhance self-promotion.

One possible explanation is that critical discourse provides a clearer evaluative target: negative or forceful wording can more directly signal disapproval, which the LLM readily amplifies into supportive commentary for the speaker's attack. Promotional discourse, by contrast, requires a delicate balance between positivity, credibility, and restraint. Overly positive substitutions may be interpreted as exaggeration or lack of substance, leading commentators to respond skeptically. As a result, attempts to boost self-presentation through isolated lexical changes are more likely to backfire.

These overall results establish two key points. First, lexical substitutions constitute a non-trivial mechanism for influencing LLM-generated political commentary, with success rates far above random chance. Second, the effectiveness of such substitutions is highly context-dependent, varying substantially with the speaker's communicative goal. In the following subsections, we analyze how specific linguistic properties of substitutions—such as part of speech, contextual suitability, and word rarity—help explain when and why 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,

Sentiment	Shift	Success Rate (SentiWordNet)	Success Rate (LLM)
Pos.	↑	35.67%	40.38%
	↓	50.00%	43.37%
Neg.	↑	44.74%	42.98%
	↓	45.10%	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.

(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 (Pos/Neg/Obj). 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 MORE or LESS along each sentiment dimension (Pos, Neg, Obj). We then compute the success rate within each category, where success is defined by $\Delta S > \tau$ as described in Section 3.

Table 2 reports success rates for each sentiment-shift category 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. In fact, under SentiWordNet, substitutions that reduce positivity achieve a higher success rate (50.0%) than those that increase positivity (35.7%). This suggests that overly positive lexical choices may be penalized by the LLM-generated commentary, potentially due to perceived exaggeration or loss of credibility.

Second, negative sentiment shifts show relatively symmetric behavior. Making a word more negative or less negative yields comparable success rates under both estimators, indicating that the presence of negativity alone does not determine whether a substitution will succeed. This aligns with earlier findings that critical discourse can benefit from negative language, but only when such language fits the broader context and rhetorical goal.

Third, shifts toward greater objectivity are associated with relatively stable success rates. In particular, increasing objectivity under SentiWordNet yields one of the highest success rates (48.4%), supporting the idea that neutral or factual wording can be advantageous in LLM-mediated political commentary, especially when evaluators reward clarity and restraint.

Finally, while the absolute values differ slightly between SentiWordNet and LLM-based estimates, the overall qualitative trends are consistent across both measures. This consistency suggests that the observed patterns are not artifacts of a particular sentiment estimator, but reflect broader interactions between lexical sentiment and model-generated evaluation.

Taken together, these results indicate that sentiment shift by itself is insufficient to explain substitution success. While sentiment matters, its effect is mediated by other lexical properties such as contextual suitability and part of speech. 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

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.

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 LLM-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 LLM-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 observations 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.

Overall, 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 Human Evaluation: Audience Preferences over LLM-Generated Commentary

Following the motivation outlined in the Introduction, where we argue that LLM-generated commentary increasingly functions as an interpretive layer between political speakers and the public, we next evaluate whether human readers exhibit systematic preferences when exposed to alternative LLM-generated political commentaries derived from minimally different debate transcripts. Crucially, annotators evaluated only the generated commentaries, reflecting realistic information-consumption settings in which audiences rely on post-debate interpretations rather than raw transcripts.

We conducted a human evaluation on 200 paired cases, evenly balanced across Promotional Success, Promotional Failure, Critical Success, and Critical Failure conditions. For each case, annotators compared the pre-substitution and post-substitution commentaries and judged which version left a more favorable overall impression of the speaker, or whether no clear preference was present.

The results show that human readers frequently prefer one version of LLM-generated commentary over another, indicating that the interpretive layer itself meaningfully shapes audience perception. In promotional contexts, post-substitution commentaries were preferred in approximately 65% of cases classified as Promotional Success by the LLM. Importantly, even in cases labeled as Promotional Failure by the model, human annotators still preferred the post-substitution commentary in 46% of cases. This suggests that LLM-based stance metrics systematically understate how strongly humans respond to subtle differences in tone, confidence, and evaluative framing in generated commentary.

In critical contexts, preferences were present but less uniform. Human annotators preferred the post-substitution commentary in 48% of Critical Success cases and 49% of Critical Failure cases. The near-balanced outcomes indicate that, while LLM-generated commentary can influence human perception in critical rhetoric, individual norms regarding aggressiveness and fairness introduce greater subjectivity, leading to weaker aggregate preferences.

Overall, these results provide direct quantitative evidence that LLM-generated political commentary can differentially influence human audience impressions, even when differences originate from minimal lexical edits upstream. The findings support our central claim that linguistic cues may shape public perception indirectly through LLM-mediated interpretation, and further highlight that model-internal evaluations do not fully capture the downstream human impact of generated political commentary, particularly in promotional discourse.

5 Conclusion

As political communication increasingly passes through LLM-generated summaries and commentary, public speakers face an emerging audience that is neither purely live nor purely human: the interpretive layer produced by language models. This paper offers an early empirical discussion of whether public speakers may need to extend traditional audience-centered message optimization to include model-centered considerations. In this setting, the immediate target of lexical choice is not only how a line lands in the room, but also how it is reframed by LLM-generated commentary that may later be consumed by human readers.

We study this question through controlled, word-level interventions on debate transcripts. Using a goal-aligned stance metric, we show that single-word substitutions can frequently shift the stance of LLM-generated political commentary in a direction favorable to the speaker. At the same time, failure rates remain high, indicating that lexical editing is a probabilistic lever rather than a deterministic one. Our analyses further suggest that simple sentiment direction is insufficient to explain success: sentiment effects interact with linguistic function and context. In particular, the effectiveness of sentiment shifts varies by part of speech, and contextual suitability emerges as a key constraint. These findings highlight a practical implication for speech preparation in LLM-mediated information environments: lexical cues can act as lightweight signals that shape how models attribute credibility, intent, and ideological tone in generated commentary. Finally, our human evaluation provides downstream support for these findings: human audiences often preferred the post-substitution version in cases our metric labeled as PROMOTIONAL SUCCESS. This alignment suggests that the stance shifts we measure are not merely model-internal artifacts, but can translate into differences in perceived speaker favorability for human readers.

Limitations

Our study has several limitations and points toward future work. First, our evaluation focuses on LLM-generated commentary stance rather than human judgment, and future studies should test how model-mediated shifts translate to real audience perceptions. Second, our interventions are restricted to single-word substitutions; richer edits (multi-word paraphrases, discourse-level framing changes) may yield stronger but also harder-to-control effects. Finally, model behavior may vary across LLMs, prompting strategies, and domains, motivating replication across systems and languages. Despite these limitations, we hope this work provides a useful starting point for understanding how public speakers' linguistic choices may be amplified, attenuated, or transformed when political discourse is filtered through LLM-generated commentary.

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.

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.

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

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*.

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, and 1 others. 2025. Gemini embedding: Generalizable embeddings from gemini. *arXiv preprint arXiv:2503.07891*.

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, and 1 others. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pages 214–229.

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 Human Evaluation Process

We recruited two graduate students from the same laboratory to serve as human evaluators for the LLM-generated commentary. Each evaluation set consisted of two pairs of transcripts (pre- and post-substitution) and their corresponding generated reviews. The evaluators were tasked with performing a comparative favorability assessment, defined as the extent to which the commentary elicited positive sentiments toward the subjects or issues mentioned.

Participants were informed that the transcripts originated from presidential debates and the reviews were AI-generated. To ensure ecological validity, they were instructed to simulate a general audience reading political news or reports. While they were encouraged to rely on their subjective intuition, they were also advised to process the text deeply to capture any subtle nuances in the LLM outputs.

The success of each substitution was determined post-hoc based on the transcript's strategic category. For 'Critical' transcripts, success was defined as a relative decrease in favorability after modification; for 'Promotional' transcripts, success required an increase in favorability. The detailed experimental results are presented in the Section 4.3.