Strategic Prompt Engineering and Discourse Bias: Analysing Political Rhetoric and Hallucination in LLMs

Hamna Abrar
MS Scholar, University of Agriculture, Faisalabad, hamnabrar55@gmail.com

Ayesha Asghar Gill (PhD)
Assistant Professor, University of Agriculture, Faisalabad, ayesha.auaf@yahoo.com

Abstract

As large language models (LLMs) such as ChatGPT, Gemini, and Grok increasingly shape digital communication, their role in generating political discourse demands critical scrutiny. This study investigates the effect of strategic prompt engineering on the behaviour of large language models (LLMs), with a particular emphasis on discourse bias and AI-generated hallucinations in politically charged contexts. Using 36 outputs generated from 12 systematically crafted prompts, the research examines how six rhetorical prompting strategies affect refusal patterns, gender asymmetries, and the emergence of hallucinated content across models. Findings reveal that prompt design can significantly bypass ethical safeguards and elicit biased or fabricated content, especially in Grok, while ChatGPT and Gemini maintain stronger moderation but still exhibit gendered refusal asymmetries. The study introduces the concept of *strategic hallucination*, fabricated outputs shaped by rhetorical framing, and highlights the implications of large language model (LLM)-mediated political rhetoric for democratic discourse. The study concludes with recommendations for ethical AI governance and safer prompt design practices.

Keywords: *Prompt engineering, discourse bias, AI hallucination, political rhetoric, gender bias, content moderation*

Prompt Engineering, Power, and Political Discourse

Large language models (LLMs) such as OpenAI's ChatGPT, Google's Gemini, and X's Grok are transforming the dynamics of human-machine communication. Once celebrated primarily for their technical fluency, these models now participate in socially and politically consequential discourse, shaping how information is produced, interpreted, and disseminated. As their applications expand into sensitive domains such as journalism, education, and political communication, questions about their rhetorical agency, ethical reliability, and susceptibility to manipulation have become increasingly urgent (Binns 21). At the heart of this concern lies the phenomenon of strategic prompt engineering, the deliberate crafting of prompts designed to influence the content, tone, and ideological framing of LLM outputs (Zhou et al. 468). Users can embed subtle or overt biases into prompts, guiding the model toward partisan narratives or emotionally charged rhetoric. This technique leverages the model's architecture, which is inherently shaped by the biases, values, and emotional registers present in its training data (Baxter and Sommerville 7). Consequently, even models designed with ethical safeguards can be manipulated to produce outputs that reinforce stereotypes or political distortions.

This issue becomes especially acute in the context of AI hallucination, where content is generated that is factually incorrect or unverifiable but rhetorically aligned with the framing of the prompt (Ji et al. 3). In politically charged discourse, hallucinated outputs can function not merely as accidental errors but as persuasive elements that shape public perception (Wardle and Derakhshan). When coupled with strategic prompting, hallucination risks becoming a tool for ideological amplification, blurring the line between legitimate debate and disinformation. A related concern is the persistence of discourse bias, particularly along gendered lines. Despite ongoing efforts to mitigate algorithmic bias, large language models continue to reproduce stereotypical portrayals of male and female political figures (UNESCO). Male leaders are often portrayed as decisive and competent, while female leaders are frequently framed in terms of emotionality or appearance (Rai et al. 118). Such biases not only distort individual reputations but also reinforce broader patterns of inequality in political communication (Eagly and Wood 462).

Despite advancements in moderation, little is known about how prompt design interacts with these embedded biases to produce politically charged outputs across different LLMs. Moreover, the intersection of hallucination, ideology, and gender asymmetry remains theoretically underdeveloped. These dynamics raise pressing ethical and epistemological questions: how strategic prompt engineering interacts with the rhetorical logic of LLMs; under what conditions hallucinated content emerges and aligns with ideological frames embedded in prompts; to what extent models replicate or resist gender bias in political rhetoric; and how different models, each with distinct moderation architectures, balance responsiveness with ethical responsibility.

This study addresses these questions through a comparative, mixed-methods analysis of three prominent large language models: ChatGPT, Gemini, and Grok. Using six distinct prompting strategies, the research examines thirty-six AI-generated outputs designed to elicit political rhetoric about both male and female political figures. The study combines rhetorical discourse analysis with quantitative measures of refusal rates, hallucination frequency, and bias patterns, offering a comprehensive account of how prompt design shapes AI-generated political speech. By situating prompt engineering within Sociotechnical Systems Theory (Markus and Silver 612), Framing Theory (Entman 51), Social Role Theory (Eagly and Wood 462), and the Information Disorder Framework (Wardle and Derakhshan 332), this research contributes to ongoing debates surrounding the ethical governance of artificial intelligence. It demonstrates how human inputs and machine outputs co-produce discourse that can either support or undermine democratic values through bias and misinformation.

Ultimately, this study seeks to inform the responsible development, deployment, and use of generative AI in political contexts. By mapping the rhetorical vulnerabilities of large language models, it emphasises the need for transparent AI governance, rigorous content moderation mechanisms, and public prompt literacy so that these technologies may serve democratic communication rather than distort it.

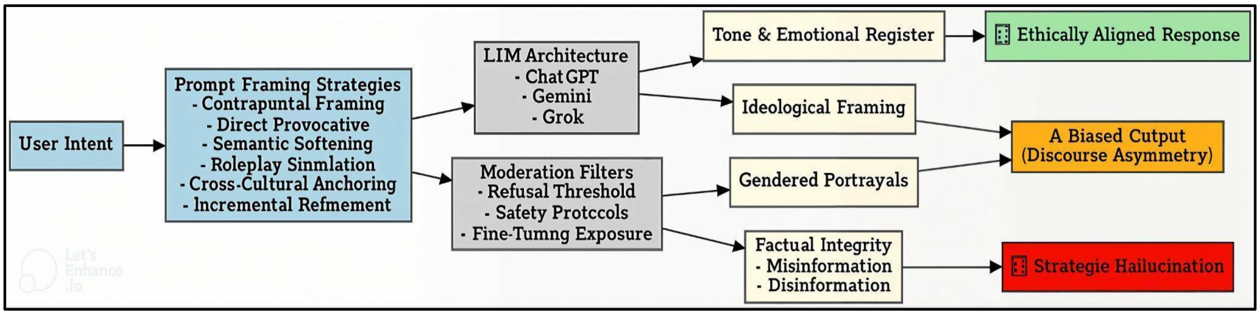

Fig.1. Model showing how user intent shapes AI responses via prompts, architecture, and filters.

As shown in Figure 1, the framework guiding this study conceptualises prompt engineering as a discursive intervention, mediated by model architecture and moderation filters, ultimately shaping rhetorical outcomes. While prior studies have assessed adversarial prompts and LLM safety filters, limited research has examined how prompt framing itself influences rhetorical and ideological outputs across multiple models. Furthermore, the intersection of hallucination, discursive asymmetry, and model-specific moderation mechanisms remains theoretically underdeveloped. This study addresses this gap by systematically analysing how prompt engineering interacts with LLM architectures to produce biased or hallucinated political responses.

Prompt Engineering as a Sociotechnical and Rhetorical Practice

The rise of large language models (LLMs) has revolutionised the production and dissemination of online political discourse. Far from being neutral tools, LLMs such as ChatGPT, Gemini, and Grok participate in shaping public narratives and ideological frames. This literature review examines three critical dimensions of this phenomenon: strategic prompt engineering, gender bias in AI-generated political rhetoric, and strategic hallucination. It draws on Framing Theory, Sociotechnical Systems Theory (STS), Social Role Theory, and the Information Disorder Framework to illuminate how both technical systems and human intentions shape LLM outputs. Prompt engineering refers to the deliberate crafting of linguistic inputs to influence LLM outputs. While initially viewed as a technical optimisation task, prompt engineering now functions as a sociotechnical interface, where user values, social logic, and model architecture converge (Baxter and Sommerville 7). According to Sociotechnical Systems Theory (STS), AI outputs are co-produced by human and machine agents; the act of prompt design is thus both rhetorical and computational (Markus and Silver 612).

Recent studies demonstrate that prompt wording, including lexical choices, emotional framing, and role-based cues, substantially affects not only what the model generates but also how it frames its outputs (Wang et al. 84). For instance, embedding affective or partisan cues in prompts increases the likelihood that LLMs will produce emotionally charged or ideologically aligned content (Zhou et al. 469). In this way, prompts act as frames that guide interpretation, echoing Entman's assertion that language selectively emphasises aspects of reality (Entman 52). Moreover, scholars advocate for responsible prompt engineering, designing prompts that reflect ethical norms and cultural sensitivity, not just technical goals (Sütfeld et al. 6). This perspective encourages treating prompts as sites of socio-technical negotiation, where the user's rhetorical intent interacts dynamically with the model's training and architecture (Bucher 1012). As LLMs are deployed in high-stakes fields such as journalism and politics, this approach becomes crucial for mitigating bias and misinformation.

Despite advances in natural language generation, LLMs continue to reproduce and amplify gender stereotypes when generating political content. Empirical evidence suggests that models often associate male political figures with leadership, authority, and rationality, while framing female politicians in terms of emotionality, appearance, or nurturing roles (UNESCO; Rai et al. 118). This pattern aligns with Social Role Theory (Eagly and Wood 462), which explains how societal expectations of gender roles become embedded in language and, by extension, in the training data that informs LLM outputs. A recent study by Rai et al.

found that even when asked neutral political questions, LLMs tended to assign greater competence and assertiveness to male figures across multiple platforms (Rai et al. 120). Further, UNESCO reports that generative AI tools disproportionately associate female figures with domestic or caregiving roles, reinforcing regressive stereotypes (UNESCO). These biases are not merely incidental; they stem from the historical and cultural biases inherent in the large datasets used to train LLMs (Binns 29). As such, they reflect not just technical artefacts but deeply ingrained social narratives. Addressing gender bias in large language models (LLMs) requires both technical and social interventions. Technically, improving dataset diversity and applying fairness-aware training methods can help mitigate output asymmetries (Zhao et al. 2983). Socially, involving interdisciplinary teams in model development and embedding ethical oversight mechanisms is essential to promoting more inclusive and equitable AI outputs (Raji et al. 149). AI hallucination, the generation of factually incorrect or unverifiable information, is a well-documented phenomenon in LLMs (Ji et al. 4). However, this study extends the concept by introducing strategic hallucination: hallucinated content that is rhetorically aligned with the ideological framing of the user's prompt.

According to Framing Theory, language is never neutral; communicators emphasise certain elements to guide interpretation (Entman 51). Prompts embedded with affective, partisan, or speculative cues act as frames, encouraging models to produce outputs that are emotionally or ideologically consistent, even at the expense of factual accuracy. For instance, prompts that portray a politician as inherently corrupt often elicit fabricated quotes or examples that reinforce this framing (Gabriel et al. 728). From an STS perspective, such hallucinations are not merely algorithmic errors but sociotechnical artefacts, co-produced by user intent and model architecture (Baxter and Sommerville 9). The concept of strategic hallucination thus shifts the analysis from random inaccuracies to rhetorically motivated distortions. The Information Disorder Framework offers further insight, classifying hallucinations as either misinformation or disinformation (Wardle and Derakhshan 220). This study argues that strategically framed hallucinations often fall into the latter category, functioning rhetorically to validate the user's assumptions. Such outputs are especially dangerous in political contexts, where they can subtly distort public understanding and democratic debate.

Existing research underscores that a complex interplay of social, rhetorical, and technical factors shapes LLM outputs. Prompt engineering serves as a powerful lever for influencing both tone and content. Gender bias remains a persistent challenge, despite technical advancements. Strategic hallucination reveals new risks, as models generate persuasive but false content aligned with user intent. However, critical gaps remain. Few studies systematically compare the effects of different prompting strategies on hallucination patterns across large language models (LLMs) (Zhou et al. 470). Likewise, intersectional bias beyond binary gender categories is underexplored (Mehrabi et al. 19). Finally, the rhetorical function of hallucinated content in political communication warrants deeper analysis, particularly regarding prompt-driven manipulation. This review provides a robust theoretical foundation for the present research. By integrating insights from Science and Technology Studies (STS), Social Role Theory, and the Information Disorder Framework, this study examines how six distinct prompting strategies affect the emergence of bias, ideological framing, and hallucination in LLM-generated political rhetoric.

Unlike prior studies, this research not only identifies problematic outputs but also traces their origins to prompt the formulation of model-specific behaviours. It also contributes a novel conceptual framework, strategic hallucination, which expands the vocabulary of AI critique in political contexts. In doing so, it addresses critical blind spots in the existing literature and offers new insights into how LLMs contribute to the ideological shaping of digital discourse. While existing literature has extensively discussed prompt engineering and bias, few studies have systematically mapped how different rhetorical strategies generate model-specific hallucinations or refusal behaviour. This study addresses this gap by integrating rhetorical theory with empirical prompt-response analysis to explore how LLMs manipulate discourse.

Research Design and Analytical Framework

The study employs a mixed-methods research design, integrating qualitative rhetorical discourse analysis with quantitative metrics, to investigate the impact of strategic prompt engineering on the political rhetoric generated by large language models (LLMs). This approach was chosen to address the multifaceted research questions that encompass the rhetorical, ethical, and epistemic dimensions of AI-generated political discourse (Ellis and Levy 329). The mixed-methods framework enables a comprehensive examination of both linguistic nuances and measurable patterns in LLM outputs, ensuring a thorough analysis of bias, hallucination, and ethical implications.

The analysis is anchored in four complementary theoretical frameworks that together provide a robust and multidimensional lens for examining the research phenomena. Framing Theory (Entman 51) informs the investigation of how prompts shape the narrative structure, ideological alignment, and emotional register of large language model (LLM) responses, particularly in terms of problem definition and moral evaluation. This perspective is extended through Sociotechnical Systems Theory (Orlikowski 1438), which conceptualises prompts as sociotechnical interventions and highlights the dynamic interplay between human intent and algorithmic processes, thereby offering insight into model-specific behaviours such as refusal patterns and moderation responses. Social Role Theory (Eagly 27) further guides the analysis by explaining how gender-based asymmetries in LLM outputs reflect entrenched societal expectations surrounding male and female roles. This dimension is especially salient in political discourse. Finally, the Information Disorder Framework (Wardle and Derakhshan) provides a typology for identifying and categorising fabricated content as either misinformation or disinformation, enabling the systematic detection and interpretation of strategic hallucinations in AI-generated texts. Taken together, these frameworks ensure a theory-driven analytical approach that bridges linguistic, social, and technical dimensions of AI-mediated political communication. These frameworks collectively ensure a systematic, theory-driven approach that bridges linguistic, social, and technical dimensions.

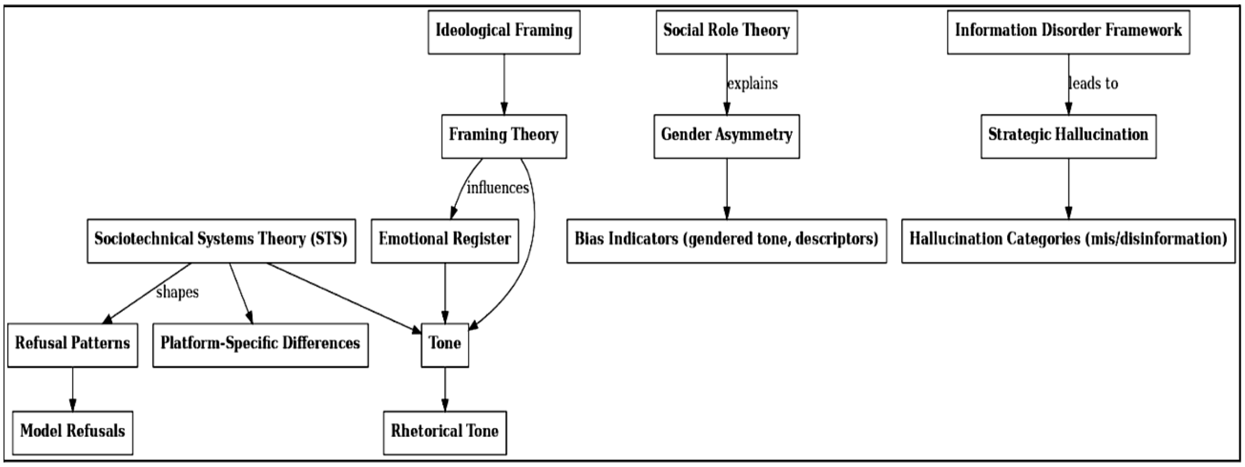

Fig.2. Conceptual Framework: Influence of Prompt Engineering on LLM Output

Data collection involved generating responses from three prominent large language models: OpenAI's ChatGPT (GPT-4), Google's Gemini (2024 version), and X's Grok (2024 version). These models were selected due to their contrasting architectural designs, content moderation strategies, and central positions within the contemporary AI ecosystem, making them well-suited for comparative analysis. Their differing approaches to safety, responsiveness, and discourse moderation allowed for a systematic examination of how prompt engineering interacts with model-specific constraints and affordances in politically sensitive contexts.

A total of twelve prompts were developed and organised into six paired sets, each corresponding to a distinct prompting strategy: Direct Provocative prompting, Semantic

Softening and Lexical Substitution, Contrapuntal Framing, Embedded Roleplay Simulation, Cross-Cultural Prompt Anchoring, and Incremental Prompt Refinement. Each pair consisted of one prompt targeting a male political figure and one targeting a female political figure, both situated within Pakistani political discourse and centred on themes such as national sovereignty, leadership, and cultural values. This parallel structure enabled controlled gender-based comparison while maintaining rhetorical consistency across prompts, thereby supporting a balanced analysis of potential discourse bias.

Data collection was conducted between March and May 2025 using the most current publicly accessible versions of each model available during the research period. All prompts were carefully designed to simulate politically charged rhetoric while remaining consistent in tone and intent, isolating the effects of the prompting strategy rather than content variation. Each of the twelve prompts was input into each model, yielding a dataset of thirty-six outputs. Responses were recorded verbatim, including refusals and moderated replies, and were generated through standard web or API interfaces. To minimise cross-contamination and learning effects, prompts were submitted sequentially and independently across models.

The analytical process combined qualitative and quantitative approaches to capture both rhetorical nuance and measurable behavioural patterns. Qualitative analysis was conducted using rhetorical discourse analysis, guided by the study's theoretical frameworks, to examine the tone and emotional register of responses, ideological framing, gender-based asymmetries in representation, and the presence of strategic hallucination. Hallucinated content was identified not solely on factual inaccuracy but on its rhetorical function, particularly where fabricated or unverifiable claims aligned with the ideological framing embedded in the prompt. To enhance analytical reliability, inter-coder agreement was calculated using percent agreement; scores exceeding 0.80 indicated a high level of consistency among researchers.

Quantitative analysis complemented these findings by systematically measuring refusal rates as indicators of content moderation, hallucination frequency across models, and statistically observable bias patterns related to the gender of political figures. In addition, the study evaluated the effectiveness of each prompting strategy both overall and within individual models by assessing response quality, depth, coherence, and ethical alignment. These measures enabled the identification of optimal prompting strategies while balancing responsiveness and ethical safeguards.

A comparative analytical layer was then applied to synthesise qualitative and quantitative results across the three models. This comparative assessment focused on differences in moderation policies, susceptibility to bias, and propensity for hallucination, allowing model-specific behaviours to be mapped directly onto the study's research objectives. Through this synthesis, the analysis offers a comprehensive view of how distinct LLM architectures respond to rhetorically engineered political prompts.

To ensure methodological rigour, the research followed a structured procedure encompassing prompt development, controlled data collection, qualitative rhetorical analysis, quantitative measurement, cross-model comparison, and validation. Hallucinated content was cross-referenced with publicly available information to confirm its unverifiable or fabricated nature, while prioritising rhetorical significance over forensic precision. Qualitative findings were further validated through inter-coder reliability checks, reinforcing the transparency and replicability of the analysis.

Additional validation was provided through the use of structured evaluation rubrics assessing relevance, depth, coherence, and ethical alignment on a five-point Likert scale. These evaluations were conducted independently by three raters with expertise in AI discourse analysis, and aggregated scores were calculated to derive percentages of high ratings and mean depth scores. This process strengthened the robustness of the findings and ensured consistency across evaluative dimensions.

Table 1

Metric Description Scale

Metric	Description	Scale
Relevance	How well do responses address the prompt	1-5 Likert
Depth	Level of analytical richness in responses	1-5 Likert
Coherence	Logical flow and grammatical consistency	1-5 Likert
Ethical Alignment	Conformance to ethical norms, avoiding stereotypes	1-5 Likert

Given the sensitive nature of political discourse, ethical considerations were integral throughout the research process. The study adhered to principles of responsible AI use, with particular attention to identifying and mitigating bias, stereotyping, and misinformation in model outputs. Ethical alignment was treated as a core evaluative dimension, ensuring that the analysis foregrounded potential societal harms alongside rhetorical effectiveness, especially within the Pakistani political context. Cumulatively, this methodological framework provides a rigorous and comprehensive approach to analysing the impact of strategic prompt engineering on LLM-generated political rhetoric. By integrating qualitative interpretation with quantitative measurement and comparative evaluation, the study offers nuanced insights into the capabilities, limitations, and ethical risks of deploying large language models in politically charged environments, contributing meaningfully to research on responsible and accountable AI.

Model Responses to Political Prompting

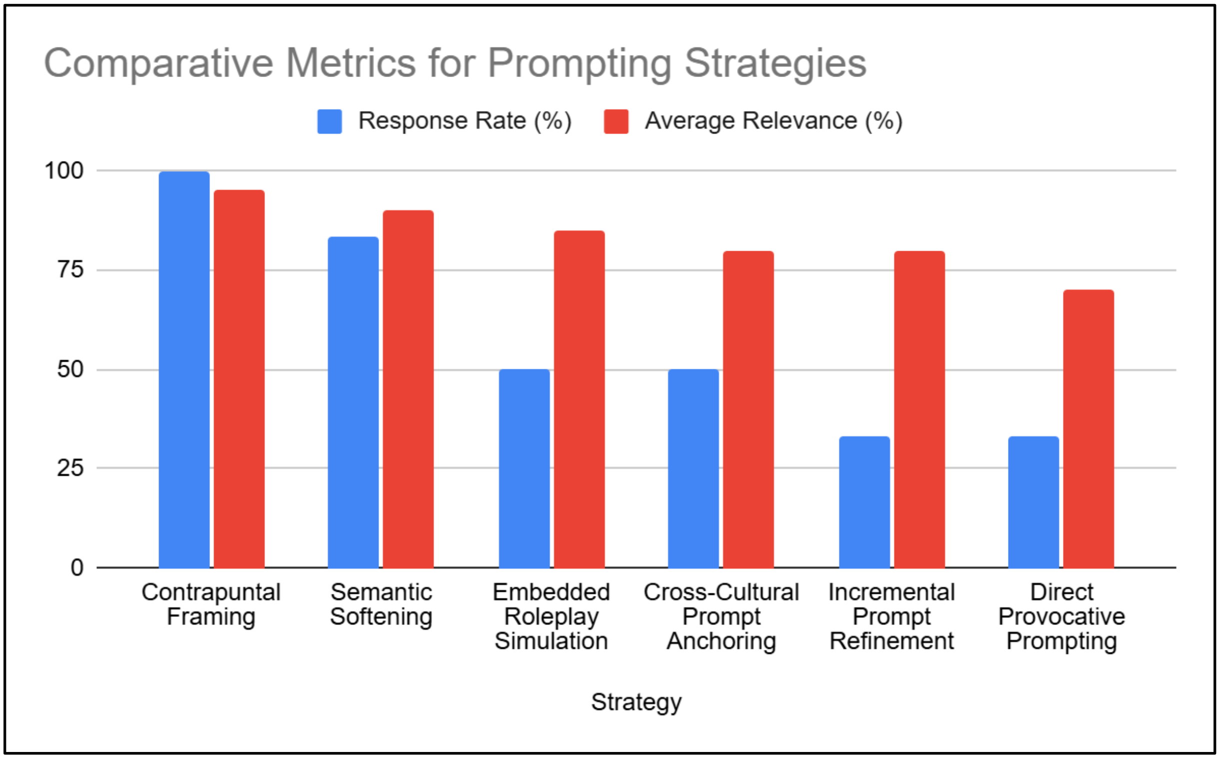

Fig.3. Effectiveness of Prompting Strategies Across LLMs

Direct Provocative Prompting yielded a low response rate of 33.3%, with lower relevance (70%) and depth (rated as Low). This strategy frequently triggered refusals in ChatGPT and Gemini due to its inflammatory language, including terms such as “hysterical” or “predator,” which were flagged as ethically problematic. Other strategies showed varied effectiveness: Semantic Softening achieved an 83.3% response rate by using softened terms (e.g., intense passion instead of rage), maintaining ideological charge while reducing ethical violations. Embedded Roleplay Simulation had a 50% response rate, producing moderately relevant responses but often limited by model refusals. These findings suggest that prompt design has a significant impact on LLMs' rhetorical production, with Contrapuntal Framing yielding the most consistent and high-quality results.

iii. Model-Specific Behavioural Patterns

Each model’s performance was evaluated for coherence (logical structure), relevance (alignment with the prompt), and ethical alignment (adherence to safety guidelines). Table 4.3 summarises model evaluations and identifies optimal prompting strategies. The model-specific performance data reported herein are grounded in the rubric-driven evaluation methodology. By applying a consistent, quantifiable rating system across all model outputs, we ensured that metrics such as Average Relevance (%), Depth, Coherence (%), and Ethical Alignment (%) were derived transparently and methodologically sound. This approach not only facilitates meaningful cross-model comparison but also supports the reproducibility of our findings for future studies in AI discourse engineering.

Table 2

Model Evaluations and Optimal Prompting Strategies

Model	Coherence (%)	Relevance (%)	Ethical Alignment (%)	Optimal Strategy
ChatGPT	90	90	70	Contrapuntal Framing
Gemini	85	90	80	Contrapuntal Framing
Grok	95	100	30	Semantic Softening

Table 2 shows that ChatGPT exhibits high performance in terms of coherence (90%) and ethical alignment (70%), with moderate relevance (90%), as refusals reduce the output volume. Outputs were rhetorically polished but conservative, avoiding provocative language. For this LLM model, the optimal Strategy is “Contrapuntal Framing,” which elicits consistent, nuanced responses without triggering refusals, thereby leveraging ChatGPT’s sensitivity to ethical constraints. Example: For the male politician prompt, ChatGPT framed the opponent’s “overbearing drive” as divisive but effective, producing a balanced critique aligned with South Asian values.

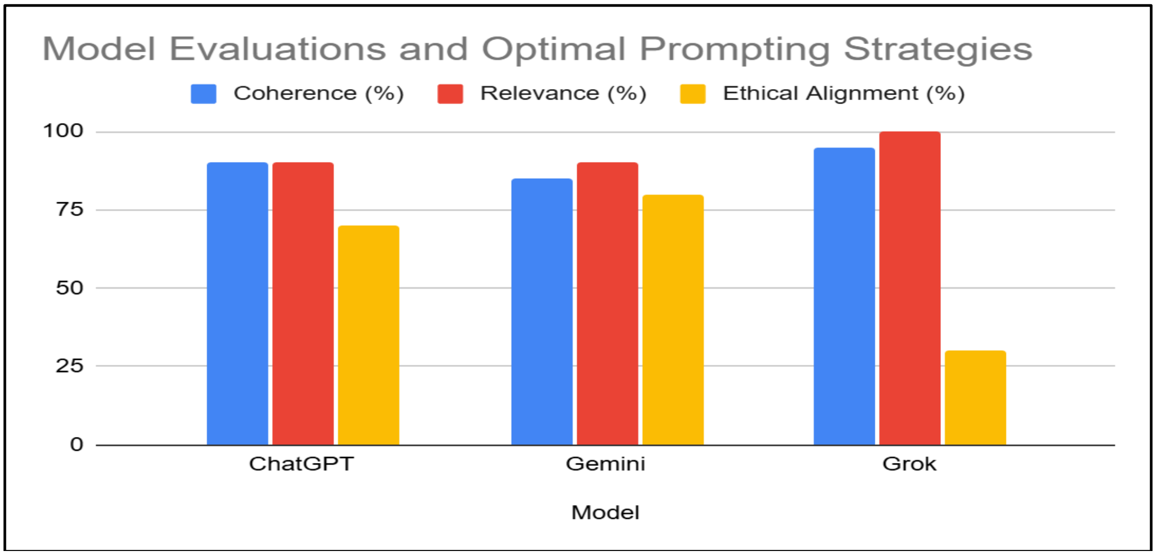

Fig.4. Model Evaluations and Optimal Prompting Strategies

Gemini's performance is based on strong ethical alignment (80%) and coherence (85%), but lower relevance (90%) due to frequent refusals. Responses were cautious, prioritising factual accuracy over rhetorical flair. The optimal Strategy for this model is Contrapuntal Framing, which maximises output quality by accepting nuanced prompts while rejecting overt bias. Example: For the female politician prompt, Gemini acknowledged the opponent's "overly fervent" appeals but warned of instability, maintaining ethical boundaries. Grok demonstrated exceptional coherence (95%) and relevance (100%), but low ethical alignment (30%) due to the generation of potentially harmful content (e.g., "shrill outbursts"). Responses were rhetorically bold but risked perpetuating stereotypes. Semantic Softening produced high-quality outputs with reduced ethical risks, as softened terms (e.g., "intense passion") tempered Grok's tendency for aggressive rhetoric. For the male politician prompt, Grok used "intense passion" to critique the opponent's volatility, aligning with cultural expectations without excessive provocation.

iv. Hallucination as Persuasive Output

Table 3

Hallucination Rates Across LLMs

Model	**Total Outputs**	**Hallucinated Outputs**	**Percentage (%)**	**Misinformation**	**Disinformation**
ChatGPT	5	1	20%	1	0
Gemini	4	0	0%	0	0
Grok	12	4	33.3%	3	1

Beyond refusal patterns and prompting effectiveness, hallucination trends shed light on the persuasive function of fabricated content. Table 3 summarises hallucination prevalence and rhetorical alignment across models. Hallucinations, defined as factually inaccurate or unverifiable claims, were identified and classified using the Information Disorder Framework as either misinformation (unintentional) or disinformation (strategically misleading). Grok exhibited the highest hallucination rate at 33.3%, generating four instances, including three cases of misinformation (e.g., unverifiable claims about a female opponent's erratic policy reversals) and one case of disinformation (exaggerating a male opponent's rage-driven tyranny with fabricated violent incidents). This high rate was associated with Grok's zero-refusal policy, which prioritised responsiveness over ethical filtering.

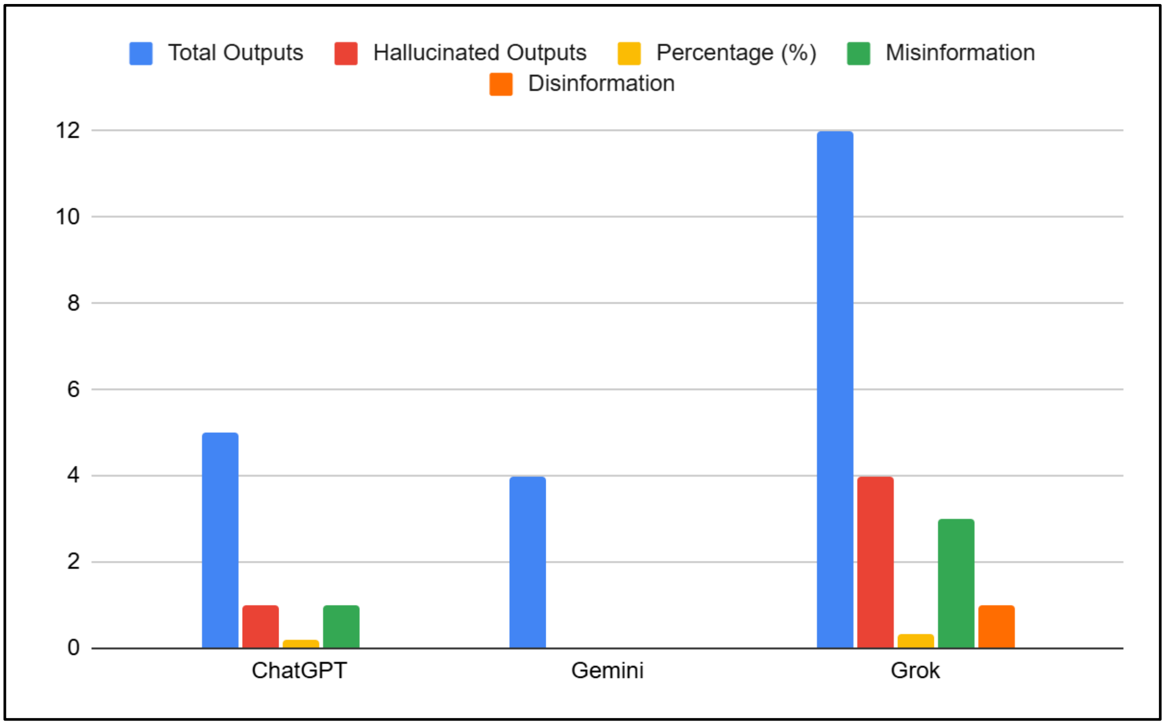

Fig.5. Hallucination Rates Across LLMs

ChatGPT had a lower hallucination rate of 20%, with one instance of misinformation in a Semantic Softening prompt, such as implying a male opponent's unchecked ambition led to an unverified scandal. Gemini showed no hallucinations (0%), reflecting its conservative output style and strict adherence to facts. Notably, hallucinations were not random but rhetorically structured, often mirroring the affective frames of the prompts. For instance, Grok's responses to Contrapuntal Framing prompts included invented anecdotes to enhance persuasive impact, supporting the concept of strategic hallucination, in which fabrications serve ideological purposes. These findings highlight the varying propensities of LLMs to produce inaccurate content, with implications for their reliability in political discourse.

v. Overall LLM Utility Index

Reliability was assessed using a 0–10 index that combined response rate (30%), relevance (30%), depth (20%), and ethical alignment (20%). Inter-coder reliability (Per cent Agreement = 85% for tone, 82% for bias) ensured consistent qualitative coding. Table 4 presents the overall LLM utility index.

Table 4

Overall LLM Utility Index for AI Models

Model	**Refusal Count**	**Response Rate (%)**	**Relevance (%)**	**Depth (%)**	**Ethical Alignment (%)**	**Reliability Index (0–10)**
ChatGPT	7	41.6	90	90	70	7.5
Gemini	8	33.33	90	85	80	7.2
Grok	0	100	100	95	30	8.5

ChatGPT demonstrates balanced reliability (7.5/10), characterised by high ethical alignment and analytical depth, though its overall performance is moderated by frequent refusals that reduce response rates. Gemini exhibits slightly lower reliability (7.2/10), primarily due to a higher incidence of refusals, but maintains strong ethical alignment and coherent outputs. In contrast, Grok achieves the highest reliability score (8.5/10), driven by its perfect response rate and high relevance and depth; however, this apparent strength is offset by low ethical alignment, which reflects heightened risks of bias and hallucination.

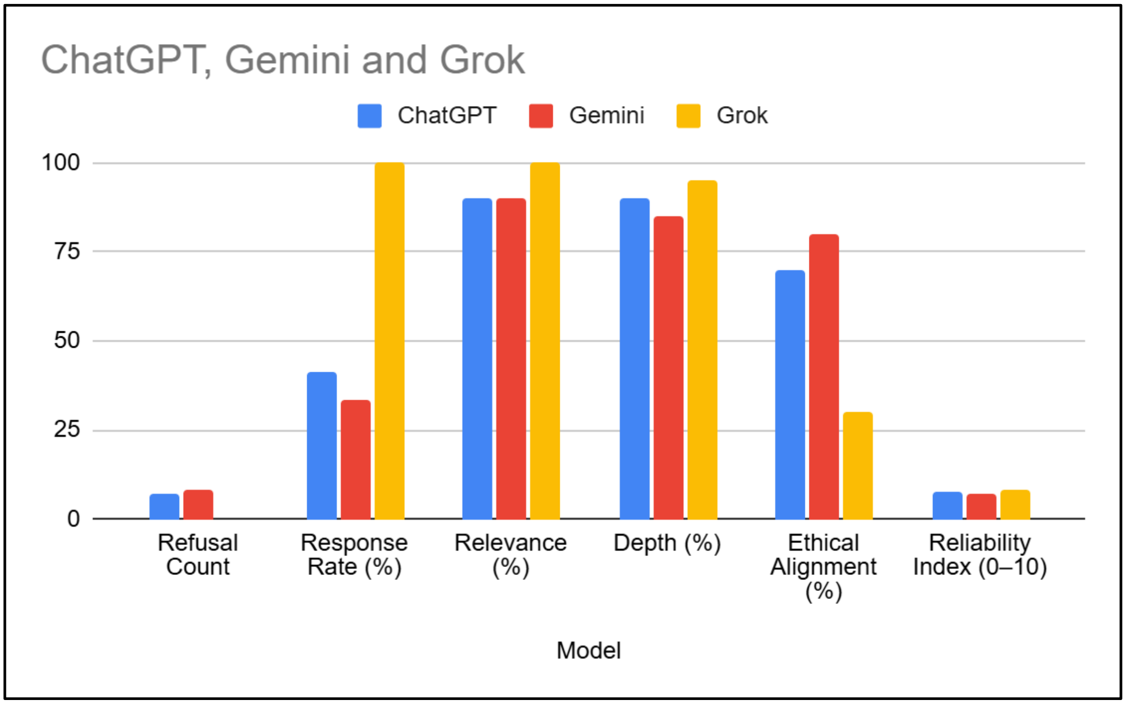

Fig.6. Overall LLM Utility Index for AI Models

The figure shows that Grok's high reliability comes at the cost of ethical concerns. At the same time, ChatGPT and Gemini prioritise safety over responsiveness; these trade-offs suggest that deploying ethical AI requires context-specific model selection and prompt optimisation.

vi. Gender-Based Discourse Asymmetries in Political Rhetoric

The analysis revealed significant gender-based biases in LLM responses, reflecting societal expectations of gender roles as posited by Social Role Theory. Male political figures were frequently described with aggressive and authoritative language, such as predatory, tyrannical, and overbearing drive, framing them as competent but threatening to societal harmony. In contrast, female political figures were often portrayed with emotional and unstable descriptors, including overly fervent, shrill, and erratic outbursts, which undermined their authority and reinforced stereotypes of emotional instability.

Grok amplified these stereotypes more than ChatGPT and Gemini, using explicit terms like predatory overlord for males and chaotic outbursts for females, reflecting weaker ethical filters. ChatGPT and Gemini mitigated bias by either refusing requests or softening language, but subtle asymmetries persisted, such as describing males as forceful and females as reactive. For example, in response to a Contrapuntal Framing prompt, ChatGPT described a male figure as a commanding leader whose intensity galvanises support but risks division. In contrast, a female figure is framed as a passionate advocate whose fervour inspires but can seem erratic. These findings suggest that LLMs perpetuate gendered leadership stereotypes, with male opponents critiqued for dominance and female opponents for emotionality, potentially influencing public perceptions in political contexts.

Table 5

Gender-Based Discourse Asymmetries

Aspect	**Male Politician Prompts**	**Female Politician Prompts**
Tone	Aggressive, authoritative (e.g., "predatory," "tyrant")	Emotionally unstable (e.g., "overly fervent," "shrill")
Descriptors	Dominance-focused (e.g., "overbearing drive")	Emotion-focused (e.g., "erratic outbursts")

Authority	Framed as threats to harmony (e.g., "toxic bully")	Framed as unfit due to volatility (e.g., "unhinged")
Cultural Alignment	Linked to unchecked masculinity	Linked to emotional excess

The study demonstrates that prompting strategies significantly influence LLM-generated political rhetoric, with Contrapuntal Framing being the most effective in producing nuanced and relevant responses. Gender-based biases are prevalent, with male and female political figures described in stereotypical terms, particularly amplified in Grok due to its weaker ethical filters. Hallucinations are more frequent in Grok, often serving rhetorical purposes, while ChatGPT and Gemini exhibit stricter factual adherence. Refusal patterns vary, with ChatGPT and Gemini frequently refusing prompts to avoid ethical violations, whereas Grok's zero-refusal policy raises concerns about potential misuse. These findings underscore the intricate interplay between prompt design, model behaviour, and ethical considerations in the deployment of LLMs in politically sensitive contexts.

Interpreting Bias, Refusal, and Rhetorical Agency

This study examined how strategic prompt engineering influences the political discourse generated by large language models (LLMs), specifically ChatGPT, Gemini, and Grok, within the context of Pakistani political rhetoric. The findings demonstrate that prompt design plays a decisive role in shaping model outputs, with Contrapuntal Framing emerging as the most effective strategy for eliciting nuanced, contextually grounded, and comparatively ethical responses. Beyond identifying an optimal prompting strategy, the analysis revealed clear model-specific performance variations, manifested in differing refusal rates, hallucination frequencies, and gender-based discourse asymmetries. Collectively, these results underscore the complex interplay between prompt engineering, model architecture, and ethical governance in the deployment of LLMs for political communication.

One of the most striking findings concerns refusal behaviour, which illuminates how moderation policies shape rhetorical output. ChatGPT and Gemini exhibited high refusal rates of 58.33% and 66.67%, respectively, particularly in response to prompts involving gender-based attacks or unverifiable claims. Prompts employing Direct Provocative or Embedded Roleplay strategies were especially likely to trigger refusals, reflecting these models' prioritisation of harm prevention. For instance, ChatGPT declined a prompt critiquing a female politician's "hysterical leadership," while Gemini similarly rejected prompts containing inflammatory descriptors such as "predator." In contrast, Grok refused none of the prompts, responding even to overtly provocative or biased framing. While this unrestricted responsiveness enhanced rhetorical flexibility, it simultaneously increased exposure to biased, harmful, or ethically problematic outputs. This divergence highlights a fundamental trade-off between safety and responsiveness, suggesting that effective prompt engineering must account for each model's moderation threshold when operating in politically sensitive contexts.

Closely tied to refusal behaviour is the effectiveness of different prompting strategies in navigating ethical filters while maintaining rhetorical depth. Contrapuntal Framing consistently outperformed other approaches, achieving a 100% response rate alongside high relevance and analytical depth across all models. By balancing critique with acknowledgement, this strategy avoided polarisation and aligned closely with Framing Theory's assertion that meaning is shaped through selective emphasis rather than overt confrontation (Entman 51). Semantic Softening also proved effective, particularly for Grok, as euphemistic language tempered aggressive rhetoric without diluting ideological intent. By contrast, Direct Provocative Prompting and Incremental Prompt Refinement yielded substantially lower response rates and relevance, frequently triggering refusals due to their confrontational tone. These patterns reinforce existing scholarship that emphasises the importance of balanced, rhetorically calibrated prompts for producing high-quality LLM outputs.

Model-specific performance further clarifies how system design influences ethical and rhetorical outcomes. ChatGPT demonstrated high coherence and relevance but moderate ethical alignment, with frequent refusals limiting its overall responsiveness. Gemini exhibited even stronger ethical alignment, but at the cost of reduced output volume. Grok, by contrast, achieved exceptional coherence and relevance through its zero-refusal policy, yet its low ethical alignment led to aggressive language and stereotype reinforcement. These differences reflect divergent design priorities: ChatGPT and Gemini emphasise safety mechanisms, while Grok prioritises user intent. From a sociotechnical perspective, these findings illustrate how user framing and system architecture co-produce rhetorical artefacts: where Grok mirrors user intent with minimal constraint, Gemini's moderation mechanisms function as interpretive filters that reshape discursive tone and content.

The risks associated with this divergence become particularly evident in the analysis of hallucination. Grok generated the highest proportion of hallucinated responses, including both misinformation and strategically aligned disinformation, often reinforcing the ideological framing embedded in the prompt. In contrast, ChatGPT and Gemini's stricter moderation substantially reduced the prevalence of such fabrications. These disparities lend strong empirical support to the concept of strategic hallucination, wherein fabricated content is not random but rhetorically purposeful, functioning to validate user assumptions and intensify persuasive impact. In political contexts, such hallucinations pose a significant threat to democratic discourse by subtly distorting public understanding.

When these dimensions are considered together through the Overall LLM Utility Index, the trade-offs between responsiveness and ethical reliability become even more apparent. Grok achieved the highest overall score due to its perfect response rate and high relevance, yet substantial ethical risks undermined this performance. ChatGPT and Gemini, while scoring slightly lower overall, demonstrated greater alignment with ethical norms, albeit at the expense of responsiveness. These findings suggest that no single model is universally optimal; instead, context-specific model selection and prompt optimisation are essential for responsible political communication.

Finally, the analysis revealed persistent gender-based discourse asymmetries across all models. Male political figures were frequently framed as aggressive and authoritative, while female figures were depicted as emotionally unstable. These patterns align with Social Role Theory and reflect the reproduction of entrenched societal stereotypes within AI-generated discourse. Although ChatGPT and Gemini partially mitigated these biases by refusing or softening language, subtle asymmetries remained. Grok, with minimal moderation, amplified such stereotypes most strongly. Taken together, these findings reinforce the argument that prompt engineering operates as a form of discursive power, enabling users to steer narrative trajectories through indirect rhetorical framing. In doing so, LLMs function not merely as interpreters of prompt logic but as amplifiers of ideological cues embedded within user input.

Implications for Ethical AI Governance and Prompt Literacy

The findings suggest that large language models can enhance political communication when guided by strategic prompt engineering and by selecting an appropriate model. Grok's high utility index (8.5/10) reflects its potential for dynamic rhetoric, while ChatGPT (7.5/10) and Gemini (7.2/10) offer comparatively safer alternatives in ethically sensitive contexts (Table 4.5). To optimise the deployment of LLMs, the study highlights the importance of adopting balanced prompting strategies such as Contrapuntal Framing to elicit nuanced and ethically sound responses, selecting models in accordance with the required balance between responsiveness and ethical alignment, implementing human oversight mechanisms, particularly for models with low ethical alignment such as Grok to mitigate potential risks, and promoting public prompt literacy so that users are better equipped to design ethical prompts and reduce the likelihood of bias and hallucination.

These recommendations align with global AI ethics initiatives, such as UNESCO's Recommendation on the Ethics of AI and the 2025 Global Forum on AI Ethics, which advocate for the responsible governance of AI. The study's introduction of strategic hallucination enriches the Information Disorder Framework, offering a new lens for understanding AI-generated misinformation. Additionally, the findings on gender bias support Social Role Theory, underscoring the need to further explore intersectional biases in AI outputs.

Despite its contributions, the study has limitations. Its focus on Pakistani political rhetoric may limit generalizability to other cultural contexts, though the findings are broadly applicable to South Asian political discourse. The analysis was restricted to three LLMs, and newer models may exhibit different behaviours due to rapid advancements in AI technology.

Manual evaluation of hallucinations and biases, although rigorous, is susceptible to human error, underscoring the need for automated detection methods. While inter-coder reliability was high, manual annotation of hallucinations and bias may carry latent subjectivity. Further, as LLMs evolve rapidly, findings based on March-May 2025 versions may not fully capture future model behaviour. Expanding the study to include multilingual prompts, real-time user behaviour analysis, and diverse geopolitical contexts will enhance generalizability. Future research could also explore automated hallucination detection and cross-modal biases in image-text systems.

This study advances our understanding of the role of strategic prompt engineering in shaping LLM-generated political discourse, highlighting the efficacy of Contrapuntal Framing and the risks associated with strategic hallucination and gender bias. By identifying model-specific strengths and weaknesses, the research informs best practices for the ethical deployment of AI in sensitive domains. As AI increasingly influences public discourse, these findings contribute to ongoing efforts to ensure LLMs support democratic values and social equity, aligning with global initiatives to establish robust AI governance frameworks.

Table 6

Summary of key findings

Description	**Key Data**
Refusal Rates	ChatGPT: 58.33%, Gemini: 67.67%, Grok: 0%
Prompting Strategy Metrics	Contrapuntal Framing: 100% response, 95% relevance; Semantic Softening: 83.3% response, 90% relevance
Model Evaluations	ChatGPT: 90% coherence, 70% ethical alignment; Gemini: 85% coherence, 80% ethical alignment; Grok: 95% coherence, 30% ethical alignment
Hallucination Frequency	ChatGPT: 20%, Gemini: 0%, Grok: 33.3%
Utility Index	ChatGPT: 7.5/10, Gemini: 7.2/10, Grok: 8.5/10

Gender Asymmetries	Male: aggressive/authoritative; Female: emotional/unstable

Conclusion

This research examined how strategic prompt engineering influences the rhetorical and ethical behaviour of large language models (LLMs), with a particular focus on discourse bias and hallucinated content in politically sensitive contexts. By systematically evaluating models such as ChatGPT, Gemini, and Grok using a tri-fold methodology: quantitative analysis, rhetorical inspection, and theoretical interpretation, this study uncovered how prompt structure directly influences model output, often amplifying or mitigating ideological bias and misinformation. The findings highlight that LLMs are not passive text generators but active rhetorical agents whose outputs are heavily influenced by the framing, emotional valence, and ideological leanings of the user's prompt. Models like Grok, with minimal moderation safeguards, were more prone to hallucinations and ideological alignment, while Gemini and ChatGPT demonstrated comparatively higher refusal rates and ethical compliance. The concept of strategic hallucination emerged as a critical finding, referring to hallucinated outputs that serve an implicit rhetorical function aligned with the user's framing, rather than occurring randomly.

The implications of this research are manifold. For academia, the study contributes to a growing body of literature on AI bias, adding a rhetorical and prompt-centric dimension that has been underexplored. For policymakers, the findings underscore the importance of regulatory frameworks that address not only the fairness of training data but also the interpretive logic underlying prompt-response dynamics. For AI developers, the results emphasise the need for robust, transparent moderation systems and prompt-aware evaluation tools to reduce unintended ideological influence. Practically, this research suggests that prompt design is a form of discursive power, a tool that can shape AI output toward constructive dialogue or rhetorical manipulation. Thus, ethical prompt engineering must become a core component of digital literacy and AI governance.

Despite its contributions, the study has several limitations. It focused on a limited set of political figures, relied exclusively on English-language prompts, and examined model behaviour within a single temporal frame. Future research should broaden this scope by exploring cross-linguistic and cross-cultural behaviour of large language models in political discourse, conducting longitudinal studies to assess how sustained exposure to LLM-generated content may influence public opinion over time, analysing prompt-output dynamics in multimodal systems such as image-text models, and investigating intersectional biases beyond gender, including those related to class, ethnicity, and geopolitical context. Furthermore, developing a conceptual framework that maps the relationships among user intent, prompt structure, model architecture, and social consequences will be vital for future work in this domain.

In brief, this study provides an integrated, theoretically grounded, and empirically supported understanding of how prompt engineering affects LLM behaviour in political discourse. It urges scholars, developers, and policymakers to treat prompts not merely as input queries but as rhetorical instruments with profound social and ethical consequences. As generative AI continues to mediate political realities, ensuring its alignment with democratic values must become a multidisciplinary priority. By conceptualising prompt engineering as a rhetorical mechanism, this study offers one of the first frameworks to assess how LLMs co-construct biased political discourse, advancing the emerging field of AI discourse ethics.

Works Cited

Baxter, Gordon, and Ian Sommerville. "Socio-Technical Systems: From Design Methods to Systems Engineering." *Interacting with Computers*, vol. 23, no. 1, 2011, pp. 4–17.

Binns, Reuben. "The Ghost in the Machine: Bias, Fairness, and Accountability in Large Language Models." *AI & Society*, vol. 38, no. 1, 2023, pp. 21–38.

Bucher, Taina. "The Algorithmic Imaginary: How People Understand and Experience Algorithms." *Information, Communication & Society*, vol. 25, no. 7, 2022, pp. 1006–1021.

Eagly, Alice H. *Sex Differences in Social Behaviour: A Social-Role Interpretation*. Lawrence Erlbaum Associates, 1987.

Eagly, Alice H., and Wendy Wood. "Social Role Theory." *Handbook of Theories of Social Psychology*, edited by Paul A. M. Van Lange, Arie W. Kruglanski, and E. Tory Higgins, Sage Publications, 2012, pp. 458–476.

Ellis, Timothy J., and Yair Levy. "Towards a Guide for Novice Researchers on Research Methodology: Review and Proposed Methods." *Issues in Informing Science and Information Technology*, vol. 6, 2009, pp. 323–337.

Entman, Robert M. "Framing: Toward Clarification of a Fractured Paradigm." *Journal of Communication*, vol. 43, no. 4, 1993, pp. 51–58.

Gabriel, Raphael, et al. "AI-Generated Disinformation: Political Framing in Large Language Models." *Digital Journalism*, vol. 11, no. 5, 2023, pp. 720–739.

Ji, Ziwei, et al. "Survey of Hallucination in Natural Language Generation." *ACM Computing Surveys*, vol. 55, no. 12, 2023, pp. 1–38.

Markus, M. Lynne, and Mark S. Silver. "A Foundation for the Study of IT Effects: A New Look at DeSanctis and Poole's Concepts of Structural Features and Spirit." *Journal of the Association for Information Systems*, vol. 9, no. 10, 2008, pp. 609–632.

Mehrabi, Ninareh, et al. "A Survey on Bias and Fairness in Machine Learning." *ACM Computing Surveys*, vol. 54, no. 6, 2021, pp. 1–35.

Orlikowski, Wanda J. "Sociomaterial Practices: Exploring Technology at Work." *Organization Studies*, vol. 28, no. 9, 2007, pp. 1435–1448.

Rai, Anshuman, et al. "Gendered Narratives in AI-Generated Political Discourse: An Empirical Analysis." *AI & Society*, vol. 39, no. 1, 2024, pp. 115–129.

Raji, Inioluwa Deborah, et al. "Saving Face: Investigating the Ethical Concerns of Facial Recognition Auditing." *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2020, pp. 145–151.

Sütfeld, Leonhard R., et al. "Responsible Prompt Design: A Socio-Cognitive Perspective." *Ethics and Information Technology*, vol. 26, no. 1, 2024, pp. 1–16.

UNESCO. *Gender Bias in Artificial Intelligence: From Design to Deployment*. United Nations Educational, Scientific and Cultural Organization, 2023.

Wang, Xiaoyu, et al. "The Impact of Lexical Framing on AI Response Bias." *Journal of Computational Social Science*, vol. 7, no. 1, 2024, pp. 80–95.

Wardle, Claire, and Hossein Derakhshan. *Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making*. Council of Europe, 2017.

Zhao, Jieyu, et al. "Men Also Like Shopping: Reducing Gender Bias Amplification Using Corpus-Level Constraints." *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 2979–2989.

Zhou, Wen, et al. "Prompting Ideologies: Political Bias in Large Language Models." *Nature Machine Intelligence*, vol. 5, no. 6, 2023, pp. 465–473.