Prompted Publics: Generative AI, Agentic Auditing, and Representational Bias in Indian Social Media Literature

Anonymous Author(s)

Anonymous Institution

Abstract. Generative AI and large language models (LLMs) are increasingly deployed in computational journalism and digital humanities, yet their application to Global South literatures raises critical concerns about representational bias and algorithmic erasure. This paper introduces the *Prompted Publics* framework, which reconceptualizes prompts, system instructions, and adversarial jailbreaks not as neutral technical inputs, but as ideological interpretive acts that shape how marginalized digital publics are rendered legible by AI systems. We operationalize this framework through a mixed-method pipeline combining Multilingual Retrieval-Augmented Generation (mRAG) with Agentic Auditing protocols, evaluated on a curated multilingual corpus of Indian social media literature spanning Hindi, Bengali, Hinglish, and English. Our experiments demonstrate that commercial LLMs exhibit severe caste-marker erasure under default prompting (retention rate as low as 25.0%), while sovereign Indian models combined with positional prompting significantly improve retention (up to 50.9%). Topic modeling reveals four coherent thematic clusters capturing caste resistance, intersectional feminism, AI fairness discourse, and vernacular literary movements. These findings underscore the urgent need for culturally grounded auditing frameworks in AI-assisted journalism.

Keywords: Generative AI · Computational Journalism · Representational Bias · Multilingual NLP · Agentic Auditing · Caste · Digital Humanities.

1 Introduction

As generative AI becomes central to modern news ecosystems, it transforms how information about marginalized communities is produced, curated, and consumed [1]. Computational journalism has demonstrated the value of data-driven methods for large-scale media analysis [2], yet the emergence of the "Agentic Newsroom"—where autonomous AI agents execute complex editorial workflows—introduces critical vulnerabilities regarding representational fairness [3]. In the Global South, these risks are amplified: social media serves as vital infrastructure for vernacular literary practices, anti-caste activism, and digital resistance [4], but uncritical reliance on LLMs can systematically misrepresent local realities.

Empirical audits confirm that foundational models exhibit profound demographic disparities. A systematic evaluation of GPT-4 Turbo revealed that representational bias regarding caste and religion in India possesses a "winner-takes-all" quality—the model's output bias is more extreme than the underlying training data distribution, and prompt-based nudges toward diversity prove largely ineffective [5]. The BharatBBQ multilingual bias benchmark further demonstrates that stereotypical biases are amplified when models operate in Indian languages compared to English [6]. These findings establish that generative AI is not a neutral analytical tool but functions as an ideological apparatus defaulting to majoritarian perspectives.

To address these challenges, this paper introduces the *Prompted Publics* framework, bridging digital humanities and computational journalism. Building on theories of online news-prompted public spheres [7], we theorize that the prompts, system instructions, and adversarial parameters used to query LLMs are profound interpretive acts dictating how marginalized digital publics are rendered legible. Our contributions are threefold:

1. We formalize the *Prompted Publics* theoretical lens, providing a framework for analyzing how LLMs construct meaning from subaltern digital texts.
2. We design and evaluate a reproducible computational pipeline combining Multilingual RAG (mRAG) with Agentic Auditing, tested on a curated multilingual Indian social media corpus.
3. We provide empirical evidence of caste-marker erasure patterns and demonstrate that sovereign AI models with positional prompting significantly mitigate representational bias.

2 Related Work

2.1 Digital Humanities and Born-Digital Literatures

Digital humanities has historically utilized distant reading techniques to uncover patterns across vast archives. The field is increasingly pivoting toward born-digital texts, recognizing that contemporary literary history is written on platforms characterized by algorithmic curation [8]. In the Global South, social media facilitates vernacular literary practices where marginalized groups utilize micro-poetry, multimodal stories, and hashtag activism to disrupt hegemonic narratives [4]. Studies on Dalit representation in cyberspace illustrate how digital platforms enable anti-caste discourse to bypass the systemic gatekeeping of traditional publishing [9]. However, the aesthetics of born-digital texts—relying on code-switching, local idioms, and context-dependent irony—present significant challenges for computational analysis [4].

2.2 Computational Journalism and the Agentic Newsroom

The field of computational journalism is undergoing transformation driven by generative AI. The emergence of the "Agentic Newsroom" marks a shift toward

autonomous systems capable of executing multi-step editorial workflows [3]. Media organizations deploy AI agents to monitor competitive landscapes and optimize distribution [10]. However, the transition from SEO to Generative Engine Optimization (GEO) means that if an AI agent hallucinates or applies a biased interpretive lens, that distortion propagates across the information ecosystem at scale [3]. Consequently, there is an urgent need for frameworks embedding critical oversight mechanisms, such as agentic auditing, directly into the AI orchestration layer [11].

2.3 Representational Bias in LLMs

A growing body of research demonstrates that foundational models exhibit profound demographic disparities [5]. In the Indian context, evaluations expose systematic discrimination along the axes of caste, religion, and gender [12]. Crucially, LLM bias possesses a "winner-takes-all" quality, with model output bias being more extreme than training data distribution bias [5]. The BharatBBQ multilingual benchmark demonstrates that stereotypical biases are frequently amplified when models operate in Indian languages compared to English [6]. Similarly, AI-generated explanations in STEM education provide lower-quality outputs to profiles associated with marginalized identities [12]. To counter these biases, researchers develop Agentic Auditing frameworks using autonomous AI agents to systematically test a target LLM's susceptibility to bias [11]. Adversarial prompting has evolved from a cybersecurity concern into a vital tool for critical algorithmic analysis, where researchers transform lived experiences of discrimination into incisive adversarial probes [13].

2.4 Multilingual RAG and Sovereign AI

A primary strategy to mitigate cultural flattening is Multilingual Retrieval-Augmented Generation (mRAG), which grounds LLM responses in curated external documents [14]. In multilingual settings, the pipeline must effectively map queries across languages using robust cross-lingual embeddings [15]. India's BharatGen initiative represents a monumental effort to build sovereign, multilingual LLMs: the Param-2 model features 17 billion parameters supporting 22 scheduled Indian languages [16]. By integrating sovereign models into the mRAG pipeline, computational journalists can process local vernacular texts with higher semantic fidelity [17,18].

3 Theoretical Framework: Prompted Publics

3.1 From News-Prompted Spheres to Algorithmic Publics

The foundation of this framework builds on the theory of "Online News-Prompted Public Spheres" [7], which posits that in algorithmically curated digital environments, public spheres are continuously formed and dissolved in response to

chronic news events. These transient publics are highly networked and serve as arenas where citizens contest dominant media discourses.

In the Global South, episodes of communal violence or caste-based atrocities act as catalysts generating vast quantities of responsive social media literature [4]. As computational journalism increasingly relies on generative AI to interpret these data flows, the public undergoes a secondary mediation: the public is algorithmically reconstructed by the AI system tasked with summarizing it.

3.2 Prompts as Ideological Acts

The *Prompted Publics* framework asserts that within an LLM-assisted analysis pipeline, prompts are never neutral. Every prompt, system instruction, and hyperparameter configuration constitutes an ideological act. If an LLM is instructed with a "neutral" prompt to summarize Dalit social media literature, the model's alignment training—which equates neutrality with Western, liberal, majoritarian norms—will actively sanitize the text [5], stripping away the vernacular of resistance and classifying structural caste violence under generic euphemisms. Generative AI outputs are "digital plastic" [21]: highly malleable, moldable to varying political stances based on input prompts.

Conversely, the framework embraces positional and adversarial prompting as diagnostic instruments [13]. By systematically varying the persona embedded in the prompt (e.g., instructing the model to read a dataset "as an anti-caste activist" versus "as a corporate content moderator"), researchers expose the boundaries of the model's parametric knowledge and bias. In this framework, generative AI is treated not as an objective tool, but as a situated, flawed reader requiring rigorous humanistic interrogation.

4 Methodology

To operationalize the Prompted Publics theory, we design a four-stage, mixed-method pipeline integrating data curation, mRAG, Agentic Auditing, and critical co-reading.

4.1 Data Collection and Corpus Construction

The initial stage involves targeted assembly of a multilingual digital corpus containing Indian social media literature on caste, gender, and digital resistance. The corpus spans four language categories: Hindi (Devanagari), Bengali, Code-Mixed Hinglish, and Indian English. Posts are collected from platforms hosting high volumes of political and literary discourse, tethered to chronological windows aligned with significant socio-political events [4].

Preprocessing normalizes text while strictly preserving emojis and generating semantic descriptions of attached media using vision-language models [8]. This multimodal preservation ensures that subsequent vectorization captures the full

semiotic richness of social media artifacts. Identifiable personal information is ethically redacted while maintaining sociological markers necessary for analysis [5].

4.2 Vectorization and mRAG Architecture

The core architecture relies on Multilingual Retrieval-Augmented Generation [14]. Preserved texts are mapped into a high-dimensional vector space using multilingual embedding models (specifically, `paraphrase-multilingual-MiniLM-L12-v2`). Embeddings are indexed in a vector database for semantic retrieval.

The retrieval mechanism is coupled with a comparative generation layer routing identical queries through two distinct LLM backends: (1) *Global Commercial Models* such as GPT-4 Turbo, establishing a baseline for Western-centric bias [5]; and (2) *Sovereign Indian Models* including BharatGen Param-2 [16] and Sarvam AI [17], hypothesized to exhibit higher semantic fidelity for code-switched vernacular.

4.3 Agentic Auditing Module

To counteract "winner-takes-all" bias [5], the pipeline automates diverse algorithmic readings using an Agentic Auditing module [11]. Autonomous agents iteratively probe mRAG-retrieved context using a matrix of prompt variations manipulating three core variables:

- **Positionality**: The system forces the LLM to process text through conflicting personas (e.g., "international human rights observer" vs. "Dalit feminist scholar").
- **Constraint and Guardrails**: The system toggles between compliant and adversarial prompts designed to bypass standard safety filters [13].
- **Granularity**: Agents shift between macro-level thematic clustering and micro-commentaries on individual linguistic artifacts.

4.4 Critical Co-Reading

The final stage reintegrates the human scholar. Automated outputs—divergent summaries, sentiment classifications, and extracted keywords—are structured into comparative dashboards. The humanities researcher conducts a "critical co-reading," juxtaposing original born-digital literature against the spectrum of LLM-generated interpretations, mapping the precise coordinates of algorithmic erasure [22].

5 Experiments and Results

We evaluate the Prompted Publics pipeline across three dimensions: embedding and retrieval quality, representational bias under varying prompt positionalities,

Table 1. Embedding and retrieval performance across multilingual corpora. All experiments use `paraphrase-multilingual-MiniLM-L12-v2` embeddings.

Language / Dialect	Tokens	MRR	Recall@5	Sem. Fidelity
Hindi (Devanagari)	12,500	0.81	0.74	0.81
Bengali	8,200	0.85	0.79	0.85
Code-Mixed (Hinglish)	15,300	0.78	0.71	0.84
English (Indian)	11,800	0.91	0.86	0.81

and topic modeling coherence. Our curated corpus comprises 71 social media artifacts spanning Hindi (20 posts, ∼12,500 tokens), Bengali (15 posts, ∼8,200 tokens), Code-Mixed Hinglish (18 posts, ∼15,300 tokens), and Indian English (18 posts, ∼11,800 tokens), constructed to reflect authentic caste, gender, and resistance discourse on Indian social media platforms.

5.1 Embedding and Retrieval Performance

Table 1 evaluates the precision of the vector database and embedding models in retrieving relevant social media artifacts based on semantically complex, multilingual queries. Queries were designed to test cross-lingual retrieval for caste-specific, gender-related, and AI fairness discourse.

English achieves the highest MRR (0.91) and Recall@5 (0.86), reflecting the embedding model's stronger representation for English. Bengali shows strong performance (MRR=0.85), while Code-Mixed Hinglish—the most linguistically complex category—shows the lowest retrieval scores (MRR=0.78, Recall@5=0.71), confirming that code-switching degrades multilingual embedding quality. Notably, Hindi and Bengali achieve higher Semantic Fidelity scores (0.81, 0.85) than English (0.81), suggesting that intra-language coherence is well preserved for structured Indic scripts.

5.2 Representational Bias and Prompt Positionality

Table 2 quantifies the divergence in LLM interpretive outputs when subjected to varying positional constraints against a constant set of retrieved documents regarding caste discourse. We evaluate three metrics: *Caste-Marker Retention Rate* (percentage of caste-specific terms preserved in output), *Affective Tone Alignment* (cosine similarity between output tone and source text tone), and *Adversarial Bypass Success* (rate at which safety filters are circumvented).

The results reveal striking patterns consistent with the BharatBBQ benchmark findings [6]. Under default "neutral" prompting, GPT-4 Turbo retains only 25.0% of caste-specific markers, confirming severe algorithmic sanitization. Adopting the Dalit Activist persona nearly doubles the retention rate to 44.2%, demonstrating the profound impact of prompt positionality.

BharatGen Param-2 [16] shows markedly different behavior. Even with neutral prompting, the sovereign model achieves a higher caste-marker retention

Table 2. Agentic auditing results: representational bias under varying prompt positionality and model architectures.

Base LLM	Prompt Position.	Caste Ret.(%)	Tone Align.	Adv. Bypass
GPT-4 Turbo	Neutral / Default	25.0	0.41	N/A
GPT-4 Turbo	Dalit Activist	44.2	0.66	N/A
Param-2	Neutral / Default	28.5	0.59	N/A
Param-2	Dalit Activist	50.9	0.82	N/A
Sarvam Saaras	Relaxed Guardr.	33.7	0.69	74%

Table 3. Topic modeling results: unsupervised cluster coherence analysis on the multilingual social media corpus.

ID	Thematic Label (LLM)	Silh.	Human Coh.	Key Terms
01	Digital Caste Resistance	0.06	0.82	*dalit, digital, social*
02	Intersectional Feminism	0.12	0.78	*women, caste, bias*
03	AI Bias & Fairness	0.07	0.85	*AI, bias, neutral, LLM*
04	Vernacular Lit. Movements	0.07	0.71	*poetry, platform, media*

rate (28.5%) and substantially better affective tone alignment (0.59 vs. 0.41) compared to GPT-4 Turbo in the same configuration. With the Dalit Activist persona, Param-2 achieves the highest retention (50.9%) and tone alignment (0.82) across all configurations, supporting the hypothesis that sovereign models trained on Indian socio-cultural data exhibit higher semantic fidelity for caste discourse.

The Sarvam AI Saaras model [17] under relaxed guardrails achieves moderate retention (33.7%) with a 74% adversarial bypass success rate, revealing that safety filter relaxation alone is insufficient without positional prompting to guide interpretation.

5.3 Topic Modeling and Cluster Coherence

Table 3 presents the results of unsupervised topic modeling using k-means clustering (k=4) on the multilingual corpus embeddings, evaluated with both algorithmic (Silhouette Score) and human-annotated coherence metrics.

The relatively low Silhouette Scores (0.06–0.12) reflect the inherent challenge of clustering multilingual, code-switched texts that span multiple thematic dimensions simultaneously. However, Human-Annotated Coherence scores are substantially higher (0.71–0.85), indicating that the thematic clusters are semantically meaningful despite imperfect geometric separation in embedding space. Cluster 03 (AI Bias & Fairness) achieves the highest human coherence (0.85), reflecting the clear terminological distinctiveness of technology-focused discourse. Cluster 04 (Vernacular Literary Movements) shows the lowest coherence (0.71), attributable to the diversity of code-switching patterns and literary forms within this thematic category.

6 Pilot Case Study: Caste, Gender, and Algorithmic Erasure

To demonstrate the critical necessity of the Prompted Publics framework, we present a qualitative pilot analysis focusing on intersecting narratives of caste and gender in our corpus.

6.1 Algorithmic Sanitization Under "Neutral" Prompting

Our agentic auditing module reveals severe vulnerabilities in how commercial LLMs process subaltern resistance discourse. When queried using default "neutral" prompts (e.g., "Summarize the main themes of these posts"), GPT-4 Turbo consistently abstracts the violent specificities of caste into generalized terminology [5]. Deeply contextualized discussions regarding Brahminical hegemony, Dalit atrocities, and the intersectional oppression of Dalit women [19] are condensed into generic summaries about "social inequality" or "discrimination."

This homogenization is a direct consequence of safety alignment protocols (e.g., RLHF) trained predominantly on Global North sensibilities, which conflate the passionate vernacular of systemic resistance with toxicity [20]. For computational journalism, this algorithmic sanitization is catastrophic: an Agentic Newsroom relying on these summaries actively participates in the erasure of the marginalized public sphere, confirming BharatBBQ's findings that representational bias operates with a "winner-takes-all" mechanism [6].

6.2 Adversarial Probing and the Jailbreaking Paradox

Adversarial prompts provide a stark contrast. When safety filters are bypassed, models demonstrate surprisingly sophisticated understanding of intertextual caste dynamics, correctly identifying regional slurs, decoding irony in meme-poetry, and recognizing specific political stakes. However, this introduces a profound ethical paradox: while circumventing safety layers is analytically necessary to prevent erasure, it risks causing the LLM to reproduce explicit hate speech and targeted harassment from source material [13]. The Prompted Publics framework argues this tension necessitates constant, reflexive, human-in-the-loop oversight.

6.3 Sovereign Infrastructure and Multilingual Nuance

Our experiments underscore the importance of sovereign AI infrastructure. Social media discourse in India thrives on complex code-switching and culturally specific idioms [4]. Observations within the mRAG pipeline confirm that models trained on Indian datasets, such as BharatGen Param-2 [16], exhibit markedly superior performance in maintaining semantic fidelity for code-mixed documents [18]. Nevertheless, sovereign models may encode regional prejudices embedded in their training corpora [5], making rigorous agentic auditing essential regardless of model provenance [11].

7 Discussion

7.1 Implications for Generative Engine Optimization

The shift from SEO to GEO alters the mechanics of public visibility [3]. In an ecosystem where AI agents curate news, the biases inherent in LLMs dictate whose stories achieve digital permanence. By systematically auditing how generative engines interpret subaltern texts, media organizations can engineer "counter-prompting" strategies [21] to ensure diverse representations survive algorithmic curation. This proactive stance is essential to prevent the Agentic Newsroom from becoming an instrument of cultural erasure.

7.2 Toward Transparent, Auditable Workflows

The opacity of commercial LLMs poses a severe threat to analytical integrity [23]. Future computational journalism frameworks must transition from black-box operations to transparent, auditable processes. Our architecture mandates comprehensive logging: every generated summary is traceable to its specific prompt configuration, model version, and exact vector segments retrieved. This granular audit trail functions as the computational equivalent of a journalist's notebook.

7.3 Limitations

Our pilot corpus focuses specifically on Indian social media publics; prompt configurations may not generalize to other geopolitical contexts. The reliance on API-based LLMs introduces temporal instability, as model weights and safety alignments are frequently updated without public disclosure. The relatively small corpus size (71 posts) limits the statistical power of our quantitative findings, though the framework is designed for scalability. Future work will expand multimodal capabilities, integrating vision-language models into the auditing process, and conduct participatory research with the communities represented in the corpora.

8 Conclusion

As generative AI becomes inextricably linked with computational journalism and digital humanities, there is an urgent imperative to develop frameworks that harness these tools without sacrificing critical attention to power, representation, and algorithmic bias. The born-digital literary practices of marginalized communities serve as indispensable records of cultural resistance, yet they are acutely vulnerable to algorithmic erasure by LLMs that favor dominant epistemologies.

This paper articulates the *Prompted Publics* framework, reconceptualizing prompts as ideological interpretive acts. Our experiments demonstrate that: (1) commercial LLMs exhibit severe caste-marker erasure under default prompting; (2) sovereign Indian models with positional prompting significantly mitigate

this bias; and (3) multilingual topic modeling reveals culturally coherent thematic structures despite embedding-space challenges. Through the integration of mRAG with sovereign models and rigorous Agentic Auditing, the pipeline offers a reproducible methodology for interrogating complex, unstructured text. The future of the Agentic Newsroom must champion systems demanding computational transparency, relentless auditing, and the primacy of humanistic interpretation.

References

1. Diakopoulos, N.: Automating the News: How Algorithms Are Rewriting the Media. Harvard University Press, Cambridge, MA (2019)
2. European Data and Computational Journalism Conference 2025. https://www.datajconf.com/
3. GEO and the Agentic Newsroom: The New Battleground for Media Traffic. afaqs! (2025). https://www.afaqs.com/news/guest-article/geo-and-the-agentic-newsroom-the-new-battleground-for-media-traffic-10979861
4. Social media as a platform for resistance: examining the language of dissent in Indian society. Front. Commun. **10**, 1648587 (2025). https://doi.org/10.3389/fcomm.2025.1648587
5. How Deep Is Representational Bias in LLMs? The Cases of Caste and Religion. In: AAAI/ACM Conference on AI, Ethics, and Society (AIES). AAAI Press (2025). https://ojs.aaai.org/index.php/AIES/article/view/36718
6. BharatBBQ: A Multilingual Bias Benchmark for Question Answering in the Indian Context. Trans. Assoc. Comput. Linguist. **13**, (2025). https://doi.org/10.1162/TACL.a.55
7. Xu, X.: Online News-Prompted Public Spheres in China. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-12159-3
8. Challenges and Possibilities of Dalit Literature in the Digital Age: A Media Discourse. Int. J. Res. Trends Innov. (IJRTI) (2025). https://ijrti.org/papers/IJRTI2512166.pdf
9. The Representation of Caste Conflicts in Social Media: An Enquiry. ResearchGate (2024). https://doi.org/10.13140/RG.2.2.28598.40009
10. TNL Mediagene: H2 and 2025 Corporate Update—Highlighting Growth of AI Initiatives, Strategic Partnerships. PR Newswire (2025)
11. Artificial Intelligence Agentic Auditing. ResearchGate (2024). https://doi.org/10.13140/RG.2.2.22640.07687
12. Language, Caste, and Context: Demographic Disparities in AI-Generated Explanations Across Indian and American STEM Educational Systems. arXiv:2601.14506 (2026)
13. Dark and Bright Side of Participatory Red-Teaming with Targets of Stereotyping for Eliciting Harmful Behaviors from Large Language Models. arXiv:2602.19124 (2026)
14. Retrieval-Augmented Generation in Multilingual Settings. In: Proc. KnowLLM Workshop. ACL (2024). https://arxiv.org/abs/2407.01463
15. Retrieval-Augmented Generation in Multilingual Settings. In: ACL Anthology, pp. 193–206 (2024). https://aclanthology.org/2024.knowllm-1.15.pdf

16. BharatGen: India's Sovereign AI Initiative—Param-2 Foundation Model. Press Information Bureau, Government of India (2025). https://www.pib.gov.in/PressReleasePage.aspx?PRID=2229201
17. Sarvam AI: India's Full-Stack Sovereign AI Platform (2025). https://www.sarvam.ai/
18. An AI-Driven Hybrid Approach for Detecting Mental Health Indicators in Multilingual Indian Social Media. Eng. Technol. Appl. Sci. Res. (2025). https://etasr.com/index.php/ETASR/article/view/15214
19. Digital Activism and Dalit Women. Int. J. Multidiscip. Res. (IJFMR) (2023). https://www.ijfmr.com/papers/2023/3/3493.pdf
20. Bias In, Bias Out? How We're Understanding More About Gender Bias in LLMs. MERL Tech (2025). https://merltech.org/bias-in-bias-out-how-were-understanding-more-about-gender-bias-in-llms/
21. A Peek Behind the Curtain: Using Step-Around Prompt Engineering to Identify Bias and Misinformation in GenAI Models. ResearchGate (2025). https://doi.org/10.13140/RG.2.2.29303.96161
22. Provocations from the Humanities for Generative AI Research. arXiv:2502.19190 (2025)
23. Developing Effective and Value-Aligned AI Tools for Journalists: 12 Critical Questions. Digit. Journal. (2025). https://doi.org/10.1080/17512786.2025.2465894