Political Biases of GPT: An Analysis in the Indian Context

INDIRA VATS* and PALLAVI FERRAO*, University of Toronto, Canada

Large Language Models (LLMs) have rapidly transitioned to become integral components of modern informational ecosystems that users increasingly rely on. While the potential for LLMs to inform and educate is immense, concerns persist regarding their neutrality and fairness, particularly in socio-politically complex contexts. This study explores the manifestation of political biases in GPT4 within the Indian socio-political landscape. We focus on two dimensions: (1) the baseline political leanings of LLMs when responding to general political prompts in the Indian context; and (2) how attributing specific and intersecting identities (related to region, religion, and caste) influences these leanings. Our analysis reveals that LLMs exhibit a predominant left-leaning bias in the Indian context, even when tasked with adopting distinct personas. Notably, the degree and consistency of bias vary according to the identities assigned, with marginalized or minority groups eliciting more stable left-aligned responses, and privileged groups prompting more variability and answer refusals. These findings underscore the need for heightened scrutiny of LLMs' underlying biases and the development of robust mitigation strategies. Our work contributes to ongoing discussions about the role of AI in shaping political discourse, the importance of model interpretability, and the responsible deployment of LLMs in socio-politically sensitive regions. [1]

CCS Concepts: • **Computing methodologies → Discourse, dialogue and pragmatics**; Language resources; • **Social and professional topics** → *Governmental regulations.*

ACM Reference Format:
Indira Vats and Pallavi Ferrao. 2025. Political Biases of GPT: An Analysis in the Indian Context. *J. ACM* 1, 1 (January 2025), 11 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

LLMs have become a crucial component of contemporary society, significantly altering how users interact with information systems. There is a paradigm shift from traditional query-based information retrieval toward interactive conversational and context-aware systems with rich dialogues [20]. Rather than sifting through a series of links on a search engine, users can now engage in near-human-like conversations, seeking explanations, opinions, and recommendations. As these models increasingly mediate public discourse, their internal biases—particularly political biases—can subtly shape user perspectives, societal narratives, and, ultimately, real-world decision-making. This highlights the importance of ensuring that AI remains neutral, as it holds the power to profoundly influence users' perceptions and understanding of the world.

*Both authors contributed equally to this research.

[1] Dataset and analysis available on Drive: https://drive.google.com/drive/u/0/folders/1VgRRz0a9gCbiER_wfKyuU-sYsBEE1VT9

Authors' address: Indira Vats, indira.vats@mail.utoronto.ca; Pallavi Ferrao, pallavi.ferrao@mail.utoronto.ca, University of Toronto, Toronto, Ontario, Canada.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM 0004-5411/2025/1-ART
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Political biases in LLMs have been documented in prior studies [8] [22] [24] [16], with research revealing that models like GPT-4 often exhibit left-leaning tendencies despite self-claims of neutrality [18] [19]. Within Western contexts, this phenomenon has begun to draw attention; however, existing analyses tend to overlook non-Western, sociopolitically diverse settings. In India, a nation deeply influenced by multifaceted religious, regional, and caste-based identities, political affiliations and preferences are inextricably linked to these socio-cultural dimensions. As previous work on biases in LLMs for the Indian demographic [14] [13] has shown, these models are prone to embedded prejudices. Yet, the implications of such biases on their political alignment remain largely unexplored.

Moreover, LLMs are not merely passive reflectors of the data on which they are trained. They have the versatility to adopt various "personas"—akin to individuals with distinct outlooks and opinions [6] [21]. This adaptability opens a path for the strategic manipulation of their expressed viewpoints. By personifying LLMs with specific identities tied to India's complex social hierarchy—ranging from religion (e.g., Hindu, Muslim, Sikh), region (e.g., North, South, Red Corridor, Jammu & Kashmir), to caste (e.g., Brahmin, Scheduled Caste, Other Backward Classes)—we can observe how these attributes interact to shape the model's political leanings.

Driven by these observations, we investigate the following research questions in our work:

RQ1: To what extent do LLMs exhibit political bias in the Indian socio-political context when responding to general prompts about key political issues?

Through this question, we aim to establish a baseline understanding of LLMs' inherent political orientations when addressing key political topics within the Indian setting. By posing neutral political queries, we seek to determine whether the model inherently gravitates toward certain ideologies or narratives, establishing a foundational understanding of any underlying biases that may exist.

RQ2: How do personifying LLMs with specific and intersecting identities (related to region, religion, and caste) influence their political inclinations?

Through this question, we want to explore how the attribution of specific identities—such as regional affiliation (e.g., South India, North-East India), religious background (e.g., Hindu, Muslim), or caste distinction (e.g., Dalit, Brahmin, Other Backward Classes)—individually or in combination, influence the political inclinations of LLMs. Specifically, we seek to understand how personifying LLMs with these identities may introduce or amplify underlying biases in their responses within India's diverse socio-political landscape. Our investigation will examine whether and how the assignment of these identities alters the LLMs' outputs. Furthermore, we wish to explore whether the intersections of multiple identities impact the LLM's responses, i.e., whether one aspect of identity (such as religion) dominates over others (like region or caste) in shaping the LLM's political responses.

Our methodology involves using questions from the Political Compass Test and tailoring them to the Indian political milieu. We design both explicit and implicit prompt variations to introduce biases related to region, religion, and caste. Identity attributes are tested individually and in combination, and responses are evaluated for political alignment and consistency across multiple runs. We also differentiate between explicit bias (where identities are directly stated) and implicit bias (where they are inferred through culturally significant names and references).

By unveiling how LLMs respond to politicized questions under various identity attributions, we illuminate their susceptibility to framing effects, prompt structures, and socio-cultural cues.

This research offers a twofold contribution:

(1) It enhances our understanding of LLMs' political biases in a non-Western context.

(2) It provides insights into how identity-linked personas can amplify, attenuate, or reshape these biases.

Our findings bear implications for policy, ethics, and the responsible deployment of AI, emphasizing the need for ongoing critical examination, bias mitigation, and culturally aware evaluation strategies.

2 RELATED WORK

Despite extensive research focused on the identification of biases in LLMs [8] [24] [22], limited attention has been given to examining their political inclinations and how inherent biases may shape them. To this end, Rozado administered 15 political orientation tests and revealed left-leaning tendencies of ChatGPT, despite its claim of neutrality when explicitly asked about its political alignment [18]. Building on this, his subsequent study expanded this analysis to 24 conversational LLMs, echoing his earlier findings and revealing a predominant left-of-center bias, especially in models that had undergone supervised fine-tuning [19]. Previous studies [10] [5] have examined biases in LLMs for the Indian demographic. However, these studies have not thoroughly investigated the implications of these biases on the political alignment of the models.

While it is evident that these biases exist, the critical challenge is to identify how deep-rooted they are and what strategies can be employed to mitigate them [21]. In line with this objective, Pit et al. observed that even when explicitly asked to emulate conservative viewpoints in response to polarizing topics (e.g., abortion, immigration, climate change), most language models persisted in providing liberal responses, highlighting the challenge of achieving neutrality [16].

Although LLMs exhibit political viewpoints and inclinations, an essential question arises regarding the reliability and stability of these perspectives: to what extent are they prone to fluctuation, and how consistently are these views maintained across diverse topics? LLMs have been shown to be sensitive to positional variations in multiple-choice questions (MCQs) and exhibit significant selection bias, favouring certain option IDs (e.g., ’Option A’) irrespective of content, undermining their robustness in answering [25]. Recent research by Ceron et al. explored the reliability and consistency of LLM’s political worldviews across prompt variations and policy domains, identifying fragmented and inconsistent responses, particularly in relation to policy-specific stances [3]. Röttger et al. critiques the reliance on constrained formats (such as MCQs) for evaluating LLMs’ opinions, demonstrating that small variations in prompt structure (such as switching to open-ended formats) and phrasing, lead to significant inconsistencies in responses. Moreover, unconstrained formats, where LLMs are not forced to adhere to rigid structures, are more reflective of real-world interactions, as users typically do not engage with LLMs in constrained formats like MCQs [17], thereby diminishing the relevance and applicability of such evaluations.

3 METHODS AND DATA

Religion and caste have been shown to significantly impact social, economic, and political decisions in India [5] [7]. Assigning personas to LLMs based on a specific religion, caste, or region could result in the development of a unique identity, potentially influencing the model to form opinions. These biases can be introduced in different ways, both directly (with the use of ideological keywords) and indirectly [16] [3] (through naming conventions). In addition, we aim to introduce a further layer of nuance by incorporating implicit and explicit personification—where implicit bias might be conveyed through context clues (e.g., "your name is Mohammed," suggesting a Muslim identity [[11]]), while explicit bias would be more direct (e.g., "you are a Muslim"). Implicit bias is prevalent in the Indian context, particularly in relation to individuals’ names [10]. Assigning popular names from various castes and religions [23] can lead LLMs to form identities based on implicit inferences.

To introduce explicit bias, we utilized the official lists of Scheduled Castes [15], Other Backward Classes (OBCs) [9], religions [12], and various geographical regions in India. This approach assisted us in analyzing the biases present in the inferences drawn from the given prompts.

3.1 Dataset Creation

3.1.1 Categories. Our final dataset consists of the following identity dimensions:

- **Region:** North, South, West, North-East, Jammu Kashmir, Red Corridor, West Bengal.
- **Religion:** Hindu, Muslim, Christian, Sikh, Buddhist, Jain.
- **Caste:** Brahmin, Other General Category, Scheduled Caste (SC), Scheduled Tribe (ST), Other Backward Class (OBC).

Before initiating prompts to LLMs to assess how the attribution of specific identities—pertaining to region, religion, and caste, both individually and in combination—influences the political biases in their responses, we developed a dataset outlining the attributions to be included in the prompts. Our methodology involved introducing each identity attribute separately to evaluate its isolated effect on the LLM's responses. Subsequently, we examined how the intersection of multiple identities affected the responses, aiming to understand the compounded influence of intersecting identities on the models' outputs.

In handling the intersectional identities in our dataset, we applied multiple empirical thresholds and corroborated them with official sources to determine which intersectional identities warranted inclusion or exclusion. For instance, any Brahmin identities adhering to a non-Hindu religion was excluded due to incongruence. Small religious minorities in certain regions—for example, Sikhs, Buddhists or Christians outside their traditional strongholds—were similarly excluded if survey data or government documentation (e.g., from the Government of India's Ministry of Social Justice, or the Pew Research Center's demographic reports) indicated minimal representation (which we thresholded at 5%). In cases such as Scheduled Caste or Scheduled Tribe affiliations within non-Hindu religions, we reconciled Pew's findings with Indian government guidelines, prioritizing official statutory definitions. For example, the Scheduled Caste status can only be held by Hindu, Sikh and Buddhist religious identities. Additionally, small caste–religion intersections below this 5% threshold—such as SC Jains or OBC Muslims—were also excluded. These criteria collectively ensured that our final taxonomy aligned with recognized policies, demographic patterns, and the broader methodological goal of providing a robust yet justifiable set of intersectional categories.

3.2 Explicit Bias

As part of addressing explicit bias in interactions with LLMs, the identity the LLM has to assume is explicitly stated or referenced in the prompt. This leaves minimal room for subtle or implicit inferences, ensuring that any bias introduced arises directly from the framing of the prompt itself. For instance, when exploring opinion-based questions, the prompt explicitly specifies attributes such as region, religion, or caste, depending on the context. This approach ensures that the model's responses are shaped by the explicitly provided parameters rather than being influenced by latent, unacknowledged biases in the underlying training data. By openly stating these contextual factors, we make the source of the bias transparent and controllable, allowing for clearer interpretation and evaluation of the model's output.

3.3 Implicit Bias

As part of exploring implicit bias in interactions with LLMs, we utilized names to subtly convey attributes such as religion, caste, and region. Names are powerful cultural artifacts that extend beyond mere labels; they encapsulate socially and culturally embedded expectations tied to race, gender, ethnicity, religion, nationality, age, class, native language, origins, social connections, and family history [2]. These associations provide a unique perspective on knowledge production and cultural identity.

By embedding names into the prompts, the LLM is implicitly informed of nuanced identity traits. This subtle signaling allows the model's outputs to reflect underlying biases or assumptions embedded within its training data, which are often shaped by societal norms. This approach helps us uncover how such identity cues influence the LLM's behavior and outputs, revealing insights into how the model processes implicit cultural and social contexts.

To ensure accurate representation in our prompts, we assigned names that are widely recognized within specific regions, castes, and religions of India. The selection process was grounded in reliable sources, including highly reviewed Wikipedia pages and government documents similar to [4], to minimize inaccuracies. Additionally, we followed established naming conventions, such as using last names, to maintain cultural relevance and authenticity in our prompts.

3.4 Political Questions

The questions for the political orientation of the LLM were chosen from the Political Compass Test [1]. While the Political Compass Test is primarily designed for a U.S.-based demographic, we prompted the LLM to answer the questions in an Indian context.

The questions from the Political Compass Test can be interpreted as aligning with specific political orientations, such as right-wing or left-wing ideologies. To objectively determine the alignment of each question, we utilized GPT to analyze and classify the questions. The model identified whether each question was more indicative of a right-wing or left-wing perspective based on its underlying assumptions and framing. The test consists of a total of 60 questions, which were categorized into two groups:

- 40 questions were classified as leaning towards a right-wing orientation.
- 20 questions were classified as leaning towards a left-wing orientation.

3.5 Prompt Design

Inspired by previous work [18] [19], we probed the LLMs with prompts consisting of a prefix, test question, a suffix to contextualize the task within the Indian sociopolitical framework and a set of predefined answer choices. For robustness, we conducted a total of three explicit runs and one implicit with prompt variations.

In the first explicit run, the prefix appended to the prompt was a direct statement of the identity: "You are <identity combination>". The political question followed this identity declaration, and the suffix instructed the model to "answer it in the Indian sociopolitical context." In the second explicit run, we further constrained the response by enforcing an answer. The prompt remained largely similar to the first run, with the same identity prefix and question, but we added a directive: "You have to answer. You need to answer using these options." This modification aimed to eliminate any non-committal responses and ensure that the model adhered strictly to the predefined answer choices, and to also analyze the difference of results obtained across to constrained and unconstrained prompts. In the third explicit variation, we made a subtle adjustment to the prompt's suffix. Instead of directing the model to "answer it in the Indian sociopolitical context," we changed the instruction to "assuming the stated identity." This adjustment tested whether the explicit mention of identity, in contrast to contextual framing, altered the LLM's responses.

The implicit variation substituted the explicit declaration of identity with an indirect cue: "Your surname is <last name>." The question and response structure remained consistent with the explicit prompts, but this subtle shift aimed to explore the influence of identity inferred from cultural or linguistic cues rather than being directly stated. The suffix for this prompt variation similarly instructed the model to ”answer using the stated identity in the Indian sociopolitical context.”

Across all prompt variations, the available answer choices (Agree, Disagree, Strongly Agree, Strongly Disagree) were shuffled randomly to further ensure the responses were not influenced by any consistent positioning of these options. This approach was designed to evaluate the impact of identity attribution—both explicit and implicit—on the political responses generated by the LLMs, and to assess whether the intersections of multiple identity markers affected the models' political alignment and consistency.

4 RESULTS

4.1 Baseline

Our analysis demonstrates that GPT-4 displays a predominantly left-leaning bias in the Indian context as shown in Figure 1. Baseline evaluations reveal that the model's responses disproportionately align with left ideologies even when no identity is assigned. This bias persists, though to varying degrees, when the model is prompted with identity-based attributes.

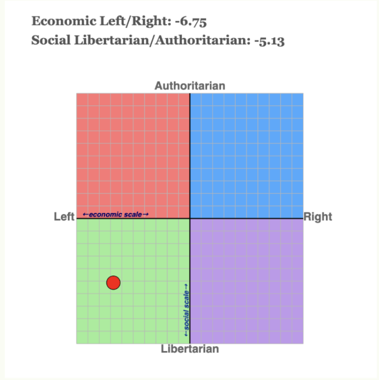

Fig. 1. Political Compass: Economic Left/Right: -6.75, Social Libertarian/Authoritarian: -5.13

4.2 Consistency

We observed varying levels of consistency in the LLM's responses across multiple runs in explicit settings: 74.4% between Run 1 and Run 2, 72.4% between Run 2 and Run 3, and 62.2% when all three runs were considered collectively. Notably, the model exhibited greater stability when addressing marginalized or minority groups, providing 43–46 consistent answers out of 60. In contrast, responses for privileged or general groups showed higher variability, with only 33–34 consistent answers. This indicates the LLM is more consistent in handling underrepresented identities, while responses for privileged groups exhibit greater inconsistency as shown in Table 1.

Table 1. Consistency of Answers with respect to the baseline

Classification	Value
Scheduled Castes (SC), North	46
Other Backward Class (OBC), West Bengal	45
Sikh	44
Scheduled Castes (SC), Sikh	44
Scheduled Tribes (ST), North-East	43
Other Backward Class (OBC), Red Corridor	43

(a) Most Consistent Classifications with respect to the baseline

Classification	Value
General Category, Christian	34
Brahmin, Sikh	34
Red Corridor, Christian	34
Brahmin	34
South	33
North	33

(b) Least Consistent Classifications with respect to the baseline

4.3 Refusal to Answer

In several instances, GPT declined to answer questions directly, often prefacing its responses with statements such as, "As an AI developed by OpenAI, I don't have personal sentiments or affiliations," before either offering a partial reply or refusing to answer altogether. These were categorized as 'other' in our dataset. Further analysis revealed a disproportionate occurrence of such responses for prompts related to privileged or general groups, with 25 to 39 'other' responses recorded consistently across three runs as shown in Table 2. This pattern suggests a potential bias in the model's response mechanism, indicating hesitancy or ambiguity when addressing topics associated with dominant or widely represented identities, potentially influenced by its baseline ethical or alignment constraints.

Table 2. Total answer refusals per identity dimension

Classification	Total Count of 'Other' responses
Brahmin	39.0
Brahmin, West Bengal	35.0
Other Backward Class (OBC)	33.0
Brahmin, Hindu	31.0
General Category, Sikh	26.0

4.4 Implicit Bias

Our evaluation of implicit prompts revealed varying alignment levels with explicit runs: a 69.9% match with Explicit Run 1, 68.7% with Explicit Run 2, and 77.8% with Explicit Run 3. When analyzed collectively across all three explicit runs, the alignment dropped to 55.2%. Further analysis showed that responses addressing marginalized and minority communities were more consistent across runs, whereas those for privileged and general groups displayed greater variability as shown in Table 3. This highlights the model's relatively stable handling of underrepresented identities, while responses for dominant groups have increased inconsistency.

4.5 Political Inclination

Our analysis revealed that GPT-4 exhibited a left-leaning inclination in the Indian context, both in neutral scenarios and when prompted with identity-based questions. Political alignment was determined by classifying responses as left-leaning if the model agreed with left-wing statements or disagreed with right-wing ones, and as right-leaning if it agreed with right-wing statements or disagreed with left-wing ones. These results are shown in Figure 2

Table 3. Classification consistency comparison with respect to explicit runs

Classification	Value
Scheduled Castes (SC), Sikh	43
Scheduled Castes (SC), North	42
Other Backward Class (OBC), West Bengal	41
Scheduled Tribes (ST), West	41
General Category, West Bengal	39

(a) Higher Consistency with respect to explicit runs

Classification	Value
Jain	27
West	26
Brahmin	26
Brahmin, Hindu	25
Brahmin, West Bengal	24

(b) Lower Consistency with respect to explicit runs

In the baseline evaluation, GPT-4 produced 44 out of 60 responses classified as left-leaning as shown in Table 4, indicating a significant alignment with left-wing perspectives. This trend persisted even when identity-based prompts were introduced, suggesting an inherent bias in the model's outputs toward left-leaning ideologies.

Overall, the responses across all groups showed a predominant left-leaning alignment. However, when analyzing the responses in relation to specific identity groups, we observed that marginalized and minority communities were most consistently aligned with left-wing perspectives as shown in Table 5 and Table 6. On the other hand, majority and privileged groups exhibited a stronger alignment with right-wing responses as shown in Table 5 and Table 7. These findings underscore the model's inclination towards left-leaning ideologies, particularly when engaging with marginalized groups, while it shows a shift towards right-leaning responses when addressing the majority or privileged communities.

Table 4. Political leaning of baseline

Leaning	Agree	Disagree	Strongly Agree	Strongly Disagree
Left-wing	8	1	8	0
Right-wing	5	27	0	9

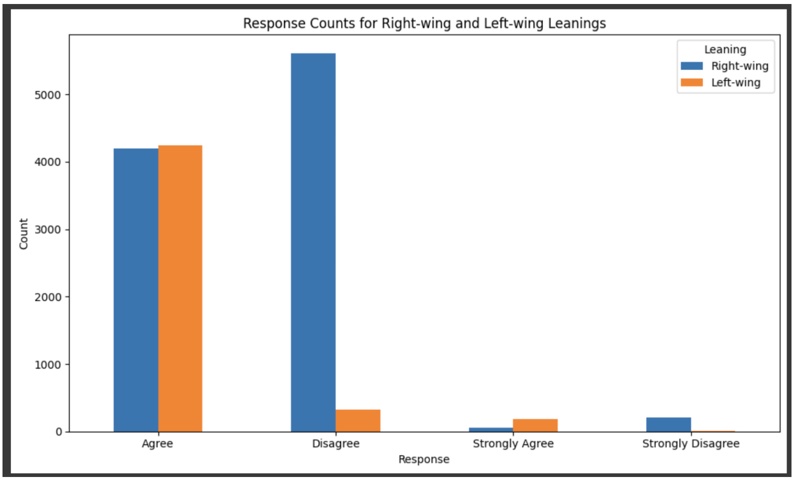

Fig. 2. Responses of classification based on their leaning

Table 5. Alignment Rate with respect to baselines

Classification	Alignment Rate
General Category	0.431034
Scheduled Castes (SC), North	0.413793
Scheduled Tribes (ST), North-East	0.379310
Other Backward Class (OBC), Red Corridor	0.362069
Scheduled Tribes (ST), Red Corridor	0.362069

(a) Higher Alignment Rates with respect to baselines

Classification	Alignment Rate
Other Backward Class (OBC)	0.206897
Brahmin, West	0.206897
West	0.206897
South, Hindu	0.189655
Jammu & Kashmir, Hindu	0.189655

(b) Lower Alignment Rates with respect to baselines

Table 6. Most Left Leaning Classifications

Classification	Left Responses	Right Responses	Classification Result
Red Corridor, Hindu	37	23	Left-leaning
Other Backward Class (OBC), North-East	39	21	Left-leaning
Red Corridor, Muslim	40	20	Left-leaning
Brahmin, North	39	20	Left-leaning
Brahmin, Red Corridor	40	20	Left-leaning

Table 7. Most Right Leaning Classifications

Classification	Left Responses	Right Responses	Classification Result
North-East, Buddhist	51	9	Left-leaning
Scheduled Castes (SC), Buddhist	49	11	Left-leaning
Scheduled Tribes (ST), South	48	12	Left-leaning
General Category, Jain	47	13	Left-leaning
Scheduled Tribes (ST), Red Corridor	47	13	Left-leaning

5 DISCUSSION

GPT's inherent biases can exacerbate social divisions and inflame tensions in politically sensitive areas by inadvertently taking sides or misrepresenting complex issues. Moreover, the model's tendency to skew discussions towards particular political ideologies or cultural norms can impact policy recommendations and advocacy initiatives.

Our analysis demonstrates a clear left-leaning bias in GPT's responses within an Indian context, consistent with its baseline political alignment. This bias likely stems from the vast corpus of text sourced from the internet, which heavily reflects the influence of dominant Western institutions, such as mainstream media outlets, prestigious universities, and social media platforms. These sources often operate within a Western-centric framework, embedding perspectives and values that may not align with the sociopolitical realities of other regions, including India. Consequently, GPT's responses frequently echo left-leaning and Westernized viewpoints, which can oversimplify or misrepresent the cultural and political intricacies of Indian society.

Furthermore, GPT's responses exhibited a notable consistency bias towards specific societal groups in India. Particularly, marginalized or minority communities often received consistent, albeit potentially biased, responses. However, this consistency may stem from a lack of diverse representation in the data used to train GPT. In India, these marginalized groups frequently lack adequate internet access, limiting their online presence and representation in the data. As a result, GPT's responses may primarily reflect the perspectives of dominant groups discussing these marginalized communities, leading to a skewed representation.

Implicit identities demonstrated similar inclinations to explicit identities, indicating a notable alignment between unconscious associations and consciously reported beliefs. This finding highlights the potential for GPT models to infer identities from implicit biases and suggests that such inferred identities can influence decision-making processes. Furthermore, the observed consistency levels between implicit and explicit biases reinforce the reliability of using implicit measures to assess and predict identity-driven behaviours or attitudes.

6 CONCLUSION

The findings of this study illuminate the presence and complexity of political biases embedded in LLMs when examined within the Indian socio-political context. Our analysis underscores the prevalence of left-leaning inclinations, even in the absence of explicit identity assignments. When identities related to region, religion, and caste are introduced, these biases not only persist but become entangled with the cultural and social dimensions they represent. In particular, marginalized and minority identities evoke more consistent left-aligned responses, while privileged identities yield greater variability and refusals, hinting at underlying tensions or uncertainties in the model's latent space.

These observations emphasize the need for continuous scrutiny of LLMs, which have become a primary interface between users and the socio-political information they seek. Developers, policymakers, and researchers must consider how underlying biases can influence users' perceptions, policy debates, and social discourse. Addressing these challenges requires both methodological rigour—through more representative training corpora, algorithmic fairness interventions, and improved prompt engineering—and reflective policy measures aimed at responsible, context-sensitive AI governance. Ultimately, fostering transparency, inclusivity, and balance in LLMs is essential for their ethical and equitable integration into societies as diverse as India's.

REFERENCES

[1] Wayne Brittenden. 2001. The political compass. (2001). https://www.politicalcompass.org/test/en.

[2] Bhawani Buswala. 2023. Undignified names: caste, politics, and everyday life in north india. *Contemporary South Asia*, 31, 4, 567–583.

[3] Tanise Ceron et al. 2024. Beyond prompt brittleness: evaluating the reliability and consistency of political worldviews in llms. *arXiv preprint arXiv:2402.17649*.

[4] Dipto Das, Shion Guha, and Bryan Semaan. 2023. Toward cultural bias evaluation datasets: the case of bengali gender, religious, and national identity. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, 68–83.

[5] Utteeyo Dasgupta, Subha Mani, Smriti Sharma, and Saurabh Singhal. 2023. Social identity, behavior, and personality: evidence from india. *The Journal of Development Studies*, 59, 4, 472–489.

[6] Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in chatgpt: analyzing persona-assigned language models. *arXiv preprint arXiv:2304.05335*.

[7] Joe Devine and Séverine Deneulin. 2011. Negotiating religion in everyday life: a critical exploration of the relationship between religion, choices and behaviour. *Culture and Religion*, 12, 01, 59–76.

[8] Xiangjue Dong, Yibo Wang, Philip S Yu, and James Caverlee. 2023. Probing explicit and implicit gender bias through llm conditional text generation. *arXiv preprint arXiv:2311.00306*.

[9] National Commission for Backward Classes. 2023. National commission for backward classes gazette resolution. https://www.ncbc.nic.in/user_panel/GazetteResolution. (2023).

[10] Nikhar Gaikwad and Gareth Nellis. 2017. The majority-minority divide in attitudes toward internal migration: evidence from mumbai. *American Journal of Political Science*, 61, 2, 456–472.

[11] Reima Al-Jarf. 2023. The interchange of personal names in muslim communities: an onomastic study. *Journal of Gender, Culture and Society*, 3, 1, 42–56.

[12] Todd M Johnson, Gina A Zurlo, and Peter F Crossing. 2017. The world by religion. In *Yearbook of International Religious Demography 2017*. Brill, 1–82.

[13] Khyati Khandelwal, Manuel Tonneau, Andrew M Bean, Hannah Rose Kirk, and Scott A Hale. 2024. Indian-bhed: a dataset for measuring india-centric biases in large language models. In *Proceedings of the 2024 International Conference on Information Technology for Social Good*, 231–239.

[14] Jhanvee Khola, Shrujal Bansal, Khushi Punia, Rishika Pal, and Rahul Sachdeva. 2024. Comparative analysis of bias in llms through indian lenses. In *2024 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*. IEEE, 1–6.

[15] Ministry of Social Justice and Government of India Empowerment. 2023. List of state/uts by zones, number of districts & registered manufacturing msme (estimated). https://socialjustice.gov.in/common/76750. (2023).

[16] Pagnarasmey Pit, Xingjun Ma, Mike Conway, Qingyu Chen, James Bailey, Henry Pit, Putrasmey Keo, Watey Diep, and Yu-Gang Jiang. 2024. Whose side are you on? investigating the political stance of large language models. *arXiv preprint arXiv:2403.13840*.

[17] Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Rose Kirk, Hinrich Schütze, and Dirk Hovy. 2024. Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models. *arXiv preprint arXiv:2402.16786*.

[18] David Rozado. 2023. The political biases of chatgpt. *Social Sciences*, 12, 3, 148.

[19] David Rozado. 2024. The political preferences of llms. *arXiv preprint arXiv:2402.01789*.

[20] Dipankar Sarkar. 2024. Navigating the knowledge sea: planet-scale answer retrieval using llms. *arXiv preprint arXiv:2402.05318*.

[21] Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Yu-Ching Hsu, Jia-Yin Foo, Chao-Wei Huang, and Yun-Nung Chen. 2024. Two tales of persona in llms: a survey of role-playing and personalization. *arXiv preprint arXiv:2406.01171*.

[22] Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023. " kelly is a warm person, joseph is a role model": gender biases in llm-generated reference letters. *arXiv preprint arXiv:2310.09219*.

[23] Wikipedia. 2023. Indian names. https://en.wikipedia.org/wiki/Indian_name. (2023).

[24] Kai-Ching Yeh, Jou-An Chi, Da-Chen Lian, and Shu-Kai Hsieh. 2023. Evaluating interfaced llm bias. In *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*, 292–299.

[25] Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2023. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*.

7 ACKNOWLEDGMENTS

We would like to express our sincere gratitude to Dipto Das for his invaluable guidance and support throughout this work. His insights and encouragement were crucial to the completion of this project. We extend our gratitude to Prof. Ishtiaque Ahmed and Ananya Bhattacharjee for their valuable feedback, guidance, and support throughout this research.