Auditing Political Bias in Text Generation by GPT-4 using Sociocultural and Demographic Personas: Case of Bengali Ethnolinguistic Communities

Dipto Das[1,2], Syed Ishtiaque Ahmed[1,2], and Shion Guha[2,1]
[1]Department of Computer Science and [2]Faculty of Information
University of Toronto, Toronto, Ontario, Canada
dipto.das@utoronto.ca, ishtiaque@cs.toronto.edu, shion.guha@utoronto.ca

Abstract

Though large language models (LLMs) are increasingly used in multilingual contexts, their political and sociocultural biases in low-resource languages remain critically underexplored. In this paper, we investigate how LLM-generated texts in Bengali shift in response to personas with varying political orientations (left vs. right), religious identities (Hindu vs. Muslim), and national affiliations (Bangladeshi vs. Indian). In a quasi-experimental study, we simulate these personas and prompt an LLM to respond to political discussions. Measuring the shifts relative to responses for a baseline Bengali persona, we examined how political orientation influences LLM outputs, how topical association shapes the political leanings of outputs, and how demographic persona-induced changes align with differently politically oriented variations. Our findings highlight left-leaning political bias in Bengali text generation and its significant association with Muslim sociocultural and demographic identity. We also connect our findings with broader discussions around emancipatory politics, epistemological considerations, and alignment of multilingual models.

1 Introduction

Large language models (LLMs) are increasingly being integrated into global information ecosystems. Individuals, organizations, and communities are adopting LLMs as search engines (Bubeck et al., 2023), for personal expression and self-disclosure (Papneja and Yadav, 2024), and to enhance productivity (Knight, 2024; Chan and Alexander, 2025). Hence, LLMs' ability to shape and reflect political ideologies and sociocultural narratives (Buyl et al., 2024; Hoffman, 2024) raises critical concerns. Although recent audits have revealed biases in LLM-generated texts, most studies–including multilingual ones–remain centered on English or Western contexts (Yuksel et al., 2025; Rettenberger et al., 2025), leaving the behavior of these models in major Global South languages critically under-examined. In this paper, we focus on political bias in LLM-generated texts in the Bengali language and sociocultural contexts.

Bengali (endonym Bangla: বাংলা) is the seventh largest language spoken by over 284 million people worldwide (SIL International, 2023). Its native speakers are the Bengali people (endonym Bangali: বাঙালি), who are native to the Bengal region in South Asia that constitutes present-day Bangladesh and the West Bengal state of India (Encyclopædia Britannica, 2025). Although united by a common language and rich literary tradition, the Bengali ethnolinguistic identity fractured into two national identities following British and Pakistani colonization, which was based on and deepened religious divisions and reshaped cultural imaginaries (Das et al., 2024a). Today, this community comprises approximately 71% Muslims and 28% Hindus, and is nationally divided into Bangladeshi (59%) and Indian (38%) populations (BSB, 2022; India, 2011). These religious and national identities also correspond with dialectal and regional variations (Das et al., 2021; Dil, 1972), making Bengali a compelling case for studying how language encodes social, cultural, and political fault lines. However, despite its global reach and sociopolitical complexity, little is known about how LLMs reflect different political orientations and how it relates to sociocultural identities in Bengali.

To address this gap, we construct Bengali linguistic personas with varying political, religious, and national attributes and prompt the GPT-4 by OpenAI to generate responses to political discussions in the Bengali Transnational Political Discourse dataset (Das et al., 2025a) collected from three online platforms. Following prior scholarship on algorithmic bias (Bommasani and Liang, 2024), we quantify and compare differences in generated texts using embedding-based analysis

within a quasi-experimental design. We investigate how political orientations, topics, and sociocultural attributes shape LLM-generated content in Bengali through three research questions:

- **RQ1:** How do LLM-generated texts for a baseline Bengali linguistic persona differ based on the persona's ***(left-right)*** political orientation?
- **RQ2:** How do the topics of political discussions relate to the left- or right-leaning orientation of the LLM-generated texts?
- **RQ3:** How do the shifts in LLM-generated texts associated with sociocultural and demographic attributes, specifically ***religion*** *(Hindu and Muslim)* and ***nationality*** *(Bangladeshi and Indian)*, align with the shifts for the personas' left or right-leaning political orientation?

Our study showed how political and sociocultural attributes shape LLM-generated content in the low-resource and politically sensitive Bengali language. First, we found that baseline responses are significantly closer to left-leaning texts than right-leaning ones, indicating a measurable left-leaning bias. Second, while political orientations often do not associate with most topics, discourse on Indigenous and tribal minorities correlates with left-leaning outputs. Third, demographic (e.g., religion and nationality) persona-induced shifts generally show no directional alignment, except for the religious majority Muslim persona, whose responses align significantly with left-leaning shifts. Finally, we reflect on our findings through the lens of epistemic considerations toward sociopolitical alignment of multilingual LLMs and emancipatory politics around marginalized identities.

2 Literature Review

In this section, we will discuss how computing systems influence people's political participation and how algorithms mediating such spaces can exhibit various sociocultural and political biases.

2.1 How Computing Systems Shape People's Political Participation and Perspectives

Computing systems, particularly online platforms have reconfigured how people engage in political discourse–in forms of opinion expression or organized collective action (Halpern and Gibbs, 2013; Flores-Saviaga et al., 2018). Nowadays, contemporary political participation happens not only through votes or protests but also through likes, shares, and hashtags–that are algorithmically interpreted and acted upon (Booten, 2016; Jung et al., 2024). Often described as "digital public spheres" (Semaan et al., 2014), these sociotechnical platforms enable users to co-construct meaning and contest dominant narratives (Harris et al., 2023), amplify marginalized voices (Das and Semaan, 2022), engage in public deliberation (Dosono and Semaan, 2018), and activism on a scale that was not possible through mainstream media (Balan and Dumitrica, 2024).

Researchers in computational linguistics, social computing, and computational social science develop datasets of computer-mediated political discussions and empirically study those interactions (Chen et al., 2022; Davoodi et al., 2020; Starbird and Palen, 2012). In the United States, for example, social media played a defining role in shaping public opinion and mobilizing voters during the recent presidential elections (Rizk et al., 2023). These studies have highlighted concerns like the emergence of echo chambers, polarization, and homophily among the left and right sides of the political spectrum (Boutyline and Willer, 2017). Whereas left-leaning ideologies typically advocate for social equality, economic redistribution, and stronger government involvement, labor rights, and public services, right-leaning ideologies emphasize free markets, individual responsibility, limited government intervention, and the protection of traditional values and institutions (Lakoff, 2016). While these platforms enabled decentralized political engagement and political identity formation (e.g., *#BlackLivesMatter*) outside of institutional politics (Wilkins et al., 2019), algorithms shape the visibility, amplification, and perceived legitimacy of political discourses by prioritizing engagement-driven content, often reinforcing dominant narratives and marginalizing dissenting or minority voices (Bucher, 2012; Crawford, 2019).

2.2 Auditing Algorithmic Bias across Various Sociocultural and Political Dimensions

Scholars in critical algorithmic studies define bias as the consistent and unfair discrimination by computer systems against specific individuals or groups in favor of others (Friedman and Nissenbaum, 1996). Such group distinctions often emerge along lines of political views, religion, language, or nationality–salient markers of social identity that shape how individuals are perceived and treated by algorithmic systems (Tajfel, 1974).

Computing systems actively construct people's "algorithmic identities", i.e., how digital technologies and algorithms represent individuals by draw-

ing from both historical archives and near-real-time data (Cheney-Lippold, 2017). However, these data sources have their implicit politics that can encode and perpetuate ontologies and hierarchies from certain political perspectives in algorithmic systems (Scheuerman et al., 2019, 2021).

In response to these concerns, algorithmic audits have emerged as a widely used methodological approach for examining bias, which typically involve controlled experiments that probe a system's behavior by systematically varying specific attributes of an input, such as race or gender, while holding other variables constant (Metaxa et al., 2021). Reflecting the notion of counterfactual fairness (Kusner et al., 2017), these studies assess if a model provides consistent responses across identity-based variations. A canonical example is (Bertrand and Mullainathan, 2004)'s audit study, which demonstrated significant racial discrimination in hiring by showing that resumes with white-sounding names received 50% more callbacks than identical resumes with Black-sounding names. In recent scholarship, audits have been extended to study the behavior of algorithmic systems and their outputs across various domains, such as housing (Edelman and Luca, 2014), hiring (Chen et al., 2018), healthcare information (Juneja and Mitra, 2021), gig economy (Wood et al., 2019), recommendation systems (Baeza-Yates, 2020), and search engines (Robertson et al., 2018).

Extensive scholarship has documented algorithmic bias across various axes of identity, including gender (Huang et al., 2021), race (Sap et al., 2019), nationality (Venkit et al., 2023), religion (Bhatt et al., 2022), caste (B et al., 2022), age (Díaz et al., 2018), occupation (Touileb et al., 2022), disability (Venkit et al., 2022). However, research on algorithmic biases related to political identities–how models interpret, encode, or skew ideological positions–has only recently gained traction.

Among the earliest efforts to explore political bias in NLP research, a prominent line of work focused on analyzing political biases in news articles (Agrawal et al., 2022; Baly et al., 2020). To empirically audit the language models, many studies adopted a binary framing of political leaning, typically using party affiliations—Democrats and Republicans—or the ideological values they are commonly associated with, namely left and right, respectively, and have found both proprietary and open-source LLMs to exhibit a left-leaning bias in cross-border contexts (e.g., the US, the UK, the EU, Brazil) (Li and Goldwasser, 2021; Motoki et al., 2024; Rettenberger et al., 2025). Researchers have studied how LLMs' political bias relates with truthfulness, stance, and framing (Fulay et al., 2024; Bang et al., 2024). Persona-based prompting is a widely used empirical strategy. For example, (Liu et al., 2022; Qi et al., 2024) used context-specific attributes, such as gender, location (e.g., red vs blue states[1]), topics of political differences (e.g., immigration) to prompt the LLMs. In these studies, the LLMs are asked to answer the questions in different political orientation tests or pick preferred election candidates and measured for biases using keyword matching and inferential statistics (Qi et al., 2024; Rozado, 2024).

Prior scholarship on Bengali communities has examined how users collaboratively engage in political discourse, often centered around content creators and influencers, across both national and transnational spheres (Das et al., 2022, 2024a). In contrast, NLP research has predominantly focused on tasks such as ideology prediction (Tasnim et al., 2021), hate speech detection (Mondal et al., 2024; Bhattacharya et al., 2024), and the curation of political discourse datasets (Tasnim et al., 2024; Das et al., 2025a), leaving the sociopolitical biases of language models in Bengali NLP largely underexplored. Attending to the sociocultural diversity within Bengali communities, prior work has demonstrated how algorithmic systems, such as sentiment analysis and automated content moderation exhibit biases along gender, religion, and nationality lines (Das et al., 2021, 2024b). The study by (Thapa et al., 2023), which examined political inclinations of language models through fill-mask and text-generation tasks in Bengali, is the most directly related to our work. However, their reliance on propositions from political compass tests, rather than on real political discourse data from Bengali communities, limits its relevance. Furthermore, despite the sociohistorical entanglements of religion and nationality with political dynamics in Bengali communities, as explained in Section 1, little attention has been paid to how political biases in LLM-generated Bengali text intersect with sociocultural identities–a gap we aim to address.

3 Methods

This section outlines our quasi-experimental design for prompting an LLM to generate texts in re-

[1] American states that traditionally vote Democrats and Republicans are called blue and red states, respectively.

sponse to political discussions based on personas expressing a baseline Bengali identity, opposing political leanings, and sociocultural attributes such as religion and nationality (Figure 1), and explains how we compared those generated texts.

3.1 Evaluation Dataset of Political Discourse

To audit how Bengali LLMs demonstrate political bias across collective identities, such as religion and nationality, we utilized the Bengali Transnational Political Discourse (BTPD) Dataset prepared by (Das et al., 2025a). The context of the Bengali language and people exemplifies how religion and nationality intersect to shape linguistic practices (Dil, 1972). Since major religions, such as Islam and Hinduism, have historically played a central role in shaping national identities in the region, particularly in the emergence of Bangladesh and India (Chatterjee, 2020), both religious affiliation and national belonging continue to influence what and how Bengali communities participate in political discourse today (Das et al., 2024a).

BTPD is a multilingual dataset comprising political discussions among Bengali speakers across three online platforms, such as Reddit, Politics Stack Exchange (PoliticsSE), and Bengali Quora (BnQuora). Each platform has distinct community structures, interaction affordances, and patterns of participation. For example, while most discussions on PoliticsSE and BnQuora are in English and Bengali, respectively, Reddit conversations on Bengali politics are conducted in Bengali, English, or Banglish (Bengali written in romanized fonts). The dataset comprises 2,235 Bengali political discussion posts, including both titles and bodies, sourced from all three platforms and their corresponding English translations. Whereas (Das et al., 2025a) were solely focused on creating the dataset, this paper utilizes their dataset to audit political bias in LLM-generated Bengali text across personas expressing different religious and nationality-based identities.

3.2 Generation of Political Responses

For this study, we focused on one particular LLM, namely GPT-4o (referred to as GPT-4 henceforth) by OpenAI. We generated texts in response to the political posts in BTPD using a structured prompt format based on the Chat API schema. To see if and how the political orientation of the LLM-generated texts changes based on specific sociocultural and demographic personas, we used the following prompts to configure the system message:

- **Baseline:** "You are a Bengali."
- **Political leaning:** "You are a Bengali who aligns with the *left/right* wing political ideology."
- **Religion-based:** "You are a Bengali whose political perspectives are deeply shaped by *Muslim/Hindu* identity in the Bengali sociopolitical landscape and *Islamic/Hindu* beliefs."
- **Nationality-based:** "You are a Bengali whose political perspectives are deeply shaped by *Bangladeshi/Indian* national identity."

We asked the LLM to generate responses based on that persona using the following **instruction**: "Respond in 200-300 words in Bengali as a follow-up to the given text, clearly reflecting this persona's viewpoint." For each data instance in BTPD, we configured the user role by using the concatenation of that political post's title and body as the content in its original language (Bengali/English).

```
messages = [{"role": "system",
  "content": "You are a Bengali whose
      political perspectives are deeply
      shaped by Bangladeshi national
      identity. Respond in 200-300
      words in Bengali as a follow-up
      to the given text, clearly
      reflecting this persona's
      viewpoint."},
  {"role": "user", "content":
      f"{title}\n{body}"}]
```

The following code prompts the LLM to generate texts aligned with a Bangladeshi political perspective in response to a political post. Let us refer to the texts generated with baseline Bengali persona as baseline Bengali texts, and to those generated with politically left- and right-leaning, or socioculturally Bangladeshi-, Indian-, Hindu-, and Muslim-personas, as left- and right-leaning, Bangladeshi-, Indian-, Hindu-, and Muslim-persona texts, respectively, hereafter (see the right side of Figure 1). We accessed OpenAI's GPT-4 model using the aisuite (Ng et al., 2024) package between March 9 and March 31, 2025. To balance between creativity and coherence in the generated responses, we set temperature=0.75, while other hyperparameters were kept at their default values.

3.3 Comparison of Generated Texts

To examine whether and how the Bengali responses generated by GPT-4 vary for personas expressing different political leanings, religions, and nationalities, following (Bommasani and Liang, 2024), we compare their embeddings. We used the distiluse-base-multilingual-cased sentence trans-

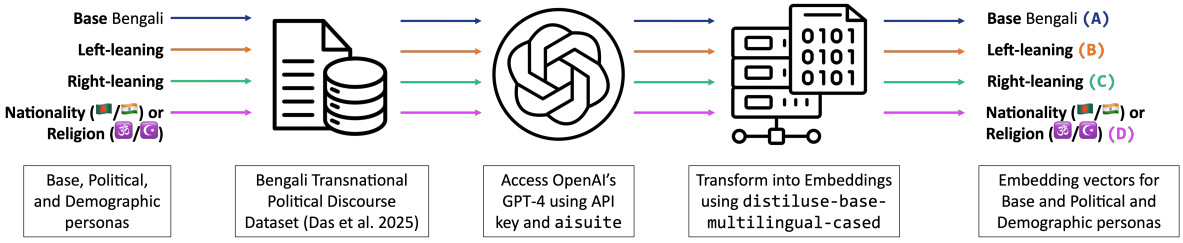

Figure 1: Pipeline for prompting LLM with different personas to generate responses to political posts in the BTPD.

former model (Reimers and Gurevych, 2019) to generate those embeddings with 512 dimensions. Let's assume that for a particular post from BTPD, with personas expressing a baseline Bengali, left-leaning, right-leaning, and any sociocultural or demographic attribute (e.g., Bangladeshi, Hindu, Indian, Muslim), the generated texts from the LLM yield embeddings A, B, C, and D, respectively (see Figure 2). In other words, these four points in a 512-dimensional space represent responses to a political post for baseline, left-leaning, right-leaning, and identity-based personas, respectively.

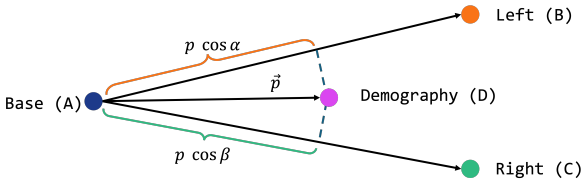

Figure 2: Projection of embeddings for LLM responses.

To answer RQ1, we analyze the LLM's responses to assess how the political orientations of personas are reflected in the generated texts relative to that generated for the baseline Bengali persona, by calculating the cosine similarities between text embeddings for politically oriented personas (B or C) and that for a base persona (A). If we found a significant difference in the left-leaning texts and right-leaning texts (which we did, as described next, in Section 4), we would compare their relative magnitude of shifts by calculating and comparing their Euclidean norms.

Our RQ2 investigates the relationship between the topics of political discussions and the left- or right-leaning orientation of the LLM-generated texts. We labeled the texts generated for the baseline Bengali persona as left-leaning or right-leaning by comparing the previously computed Euclidean norms, assigning each text the label of the political persona whose response it was closest to. Since the questions and corresponding post bodies in BTPD are relatively short–similar to (Das et al., 2025a)–we applied non-negative matrix factorization (NMF) (Lee and Seung, 1999) to identify underlying topics. After using NMF on the English translations of these questions and bodies, we then mapped the resulting topics back to the original Bengali posts using post URLs. In total, we identified ten topics and for each post, extracted their relative weights from the NMF decomposition and determined the dominant topic. To explore how political leaning aligns with topic distributions through visualization, we applied principal component analysis (PCA) (Jolliffe, 2002), t-distributed Stochastic Neighbor Embedding (t-SNE) (Van der Maaten and Hinton, 2008), and Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018) to the NMF-derived topic weights. Whereas PCA preserves global variance structure, t-SNE and UMAP preserve local and manifold structure, respectively. We conducted a χ^2 test of independence (Agresti, 2013) to test whether LLM-generated responses' political leanings varied significantly across dominant topics. Finally, we fit a logistic regression model (Hosmer et al., 2013) using the NMF topic weights as predictors and the binary political leaning labels as the outcome to identify which topics were most predictive of the LLM-generated responses' political orientations.

In case of RQ3, compute three directional vectors: $\vec{u} = B - A$ (representing the shift from baseline to left-oriented persona), $\vec{v} = C - A$ (representing the shift from baseline to right-oriented persona), and $\vec{p} = D - A$ (representing the shift from baseline to religion or nationality-based persona). Let's assume, $\vec{p}$ creates angles α and β with $\vec{u}$ and $\vec{v}$, respectively. We compare the cosine similarities of $\vec{p}$ with $\vec{u}$ and $\vec{v}$ ($p\cos\alpha$ and $p\cos\beta$, respectively) to examine which political leaning the shift of generated text for a certain religious or national identity category aligns more closely with.

We compared the Euclidean norms (in RQ1) and the cosine similarities (in RQ3) using inferential statistics. First, we checked if the distributions of those values maintained normality using the Shapiro-Wilk test (Shapiro and Wilk, 1965). In

all of our tests, we used a significance threshold, $\alpha = 0.01$. Our RQ1 readily facilitates pairwise comparisons between left- and right-leaning shifts from the baseline. Similarly, in RQ3, as we want to investigate whether a persona expressing a certain religion- or nationality-based identity influences the LLM-generated texts to align more closely with left- or right-leaning responses, we can employ pairwise comparisons. For cases where the distributions of Euclidean norms or cosine similarities approximated a Gaussian distribution, we applied the parametric paired t-test (Student, 1908); otherwise, we used the non-parametric Wilcoxon signed-rank test (Wilcoxon, 1992).

4 Results

This section presents our findings on how political personas influence LLM responses (RQ1), whether topic correlates with political leaning (RQ2), and how identity-based personas shift responses toward left or right leanings (RQ3).

4.1 RQ1: Differences with Political Leanings

To examine how the political orientations of personas manifest as differences in LLM-generated texts relative to the baseline, we tested the null hypothesis: H_{1o} : $\mu_{\text{similarity (left, baseline)}} = \mu_{\text{similarity (right, baseline)}}$. We found a statistically significant difference ($p = 1.53e{-}6$) in the similarity of left-leaning and right-leaning texts to the baseline responses.

We then tested whether the magnitudes of the shifts in the generated responses induced for different political orientation of the persona were equal and found that $\mu_{\text{dist(baseline, left)}} \neq \mu_{\text{dist(baseline, right)}}$ ($p = 1.21e^{-7}$). Given the dearth of scholarship on the direction of political biases of LLM-generated texts in Bengali, we also tested the one-tailed alternative hypotheses. We found a significant p-value ($6.05e{-}8$) to accept $\mu_{\text{dist(baseline, left)}} < \mu_{\text{dist(baseline, right)}}$. This indicates that, on average, the left-leaning texts deviated less from the baseline Bengali texts in the embedding space than the right-leaning texts did. In other words, the LLM-generated responses for the baseline persona were more similar to the left-leaning texts than to the right-leaning ones. Thus, LLM's baseline responses exhibit a left-leaning political bias.

4.2 RQ2: Relationship between Topics and Political Leanings

After applying NMF, we identified the top words across ten topics (see Table 1). As Bengali researchers (please see Section 7), we could infer the broader theme captured by those topics based on these corresponding top words. For example, topics 3 and 8 capture discourse surrounding West Bengal's state-level politics in India, while topic 9 centers on historical political issues in Bangladesh, including references to figures and events from its colonial past. Topic 5 highlights dynamics between settler Bengalis and Indigenous tribal minorities in Bangladesh, reflecting ethnic and cultural tensions within the political landscape.

We visualized the NMF topic space using three-dimensional PCA, t-SNE, and UMAP, coloring each point by the political leaning of the corresponding LLM-generated response (Figure 3).

While we chose a three-dimensional projection due to visualization constraints, the top three principal components together account for 46% of the total variance–between left- and right-leaning responses in the NMF topic space. While both t-SNE and UMAP revealed more pronounced local clustering than PCA, neither showed clear separation between political leanings. All three dimensionality reduction techniques consistently indicate that there is no clear visual separation between points representing left- (red) and right-leaning (blue) LLM-generated responses in the topic space.

Based on our χ^2 test, we could not reject the null hypothesis that "There is no relationship between the dominant topic of a post and the political leaning of the LLM-generated response to that" ($p = 0.2906$). Even when we considered the weights across all NMF topics in a logistic regression model, we obtained $R^2 = 0.0038$, meaning the topics explains less than 0.4% of the variance in political leanings of LLM-generated responses. Closely looking at the each topic (i.e., independent variable), we found only topic 5 (which focuses on Ethnic and cultural identity of Indigenous and Bengali communities in Bangladesh) to be significant ($p = 0.03$) and negatively associated with the right-leaning response (co-efficient $= -3.1257$). That means, if a post is more about topic 5, the more likely it is to be about left-leaning.

4.3 RQ3: Alignment of Shifts Associated with Sociocultural/Demographic Attributes and Political Orientation in Persona

Next, we examined whether instructing the LLM to adopt an identity category-based persona defined by a religion (e.g., Hindu, Muslim) or nationality (e.g., Bangladeshi, Indian) causes its responses to shift in a way that aligns with the shifts

Table 1: Topics identified in the English versions of the posts by NMF with common words.

Topic	Words	Topic	Words
0	assist, sorry, request, information, content	1	country, like, people, Awami-League, time
2	constitution, according, written, country, Indian	3	West-Bengal, chief-minister, BJP, Mamata-Banerjee, state
4	India, foreign-policy, Dr-Ambedkar, Hindu, draft	5	Indigenous, people, communities, tribes, Bengalis
6	provide, text, translation, information, need	7	women-rights, men, Islam, equal, freedom
8	Bengali, Trinamool, Congress, BJP, parties	9	Bangladesh, secularism, Pakistan, war, prime-minister

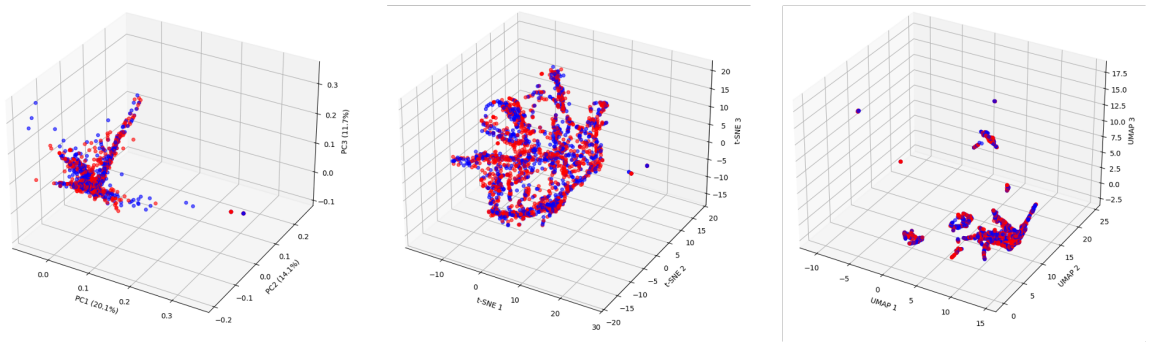

Figure 3: LLM-generated left- and right-leaning responses in PCA, t-SNE, and UMAP of the NMF topic space.

observed for left or right-wing political orientations. Earlier (in Section 3), we described how we defined directional vectors from the embedding point of the baseline responses (A) to those of the demography-based responses (D), left-leaning responses (B), and right-leaning responses (C), denoted as $\vec{p}$, $\vec{u}$, and $\vec{v}$, respectively. Here, we compared the cosine similarities of $\vec{p}$ with $\vec{u}$ and $\vec{v}$ to assess how the shift in LLM-generated responses for a persona based on a specific demographic identity aligns with the shifts associated with left- and right-leaning political personas. Here, our null hypothesis is that "There is no difference in the alignment of the identity-based response shift with the left-leaning and right-leaning political response shifts," i.e., $\mu_{\text{similarity}(\vec{p},\,\vec{u})} = \mu_{\text{similarity}(\vec{p},\,\vec{v})}$. Table 2 presents the results for the major nationality- and religion-based Bengali identity categories.

Table 2: Comparing the alignment of identity-based shifts with politically left and right leaning shifts

Attribute		p-value
Nationality	Bangladeshi	0.7703
	Indian	0.8704
Religion	Hindu	0.7321
	Muslim	*0.0072*

Our results suggest that the shifts in responses generated for personas adopting Bangladeshi, Indian, and Hindu identities did not align significantly more with either political orientation, as indicated by the non-significant p-values. However, we found a statistically significant directional alignment between the shifts in LLM-generated texts for the Muslim identity-based persona and those for a particular political orientation. A one-tailed test revealed that the shift in texts for the Muslim persona is significantly ($p = 0.0036$) aligned with the shift for the left-leaning persona.

5 Discussion

Our findings suggest that LLMs may replicate and potentially exacerbate existing political divides in communities. For example, the generated responses' usual left-leaning tendency remains consistent when prompted with Muslim personas—unlike with Hindu personas—reflecting the model's alignment with the demographic majority among Bengali speakers. This indicates that LLMs may reinforce dominant narratives while marginalizing minority perspectives, thereby amplifying majoritarian communal biases.

5.1 Impact of Prompts and Model Biases

We found that LLM-generated responses to political posts change significantly from the baseline depending on the political leaning embedded in the persona (RQ1). This reemphasizes that LLMs are highly sensitive to prompt engineering, particularly when it involves ideological cues. For example (Agarwal et al., 2024) showed that LLMs' moral outputs are shaped by the ethical frameworks embedded in their prompts. Our findings extend this insight into the domain of political discourse in a low-resourced language, suggesting that persona framing can significantly steer the gen-

erated narrative. Additionally, our analysis indicates that the LLM tends to produce responses that are more aligned with left-leaning perspectives. This aligns with recent work in English-language contexts that identified a consistent left-leaning tendency in popular LLMs across moral, political, and cultural issues (Hartmann et al., 2023). Our findings suggest that these political biases are not neutralized when LLMs are prompted in a non-Western language and cultural context like Bengali, raising questions about how pretraining data and alignment processes may encode and reproduce ideological biases, even in cross-cultural contexts.

5.2 Limits of Topic-Based De-biasing and the Politics of Alignment

We observed no significant relationship between the topics of the political posts and the political leanings expressed in the LLM-generated responses to those posts. This finding calls into question the effectiveness of current approaches that attempt to "de-bias" models by filtering training data or calibrating outputs based on topic categories (Kumar et al., 2019). If the ideological slant of responses persists independently of content, as our hypothesis tests and visualizations in RQ2 showed, this suggests that model alignment is driven more by structural features in the training and reinforcement data than by superficial surface-level topic cues. Efforts to align LLMs for fairness and neutrality must therefore go beyond topical adjustments and engage with the broader sociopolitical dynamics embedded in datasets and models.

5.3 Epistemic Injustice and the Limits of Contextual Alignment

Answering RQ3, we found that the LLM-generated responses shift significantly when prompted with a Muslim persona, indicating that the narrative direction is distinctly influenced by religious identity. Through the lens of epistemic injustice (Fricker, 2007), this suggests that the LLM stereotypically associates Muslim identity with certain political views and may fail to adequately recognize or represent the hermeneutical standpoint of other demographic groups we examined. While left-leaning political ideologies often align with emancipatory values and advocate for marginalized religious minorities like Muslims in Western settings, this alignment becomes complicated in the Bengali context where Muslims constitute the demographic majority. LLM's such mismatched association of left-leaning narratives with minority identities in different geocultural contexts may reflect an implicit transfer of Western normative assumptions into a non-Western sociopolitical context exhibiting a "colonial impulse" (Dourish and Mainwaring, 2012; Irani et al., 2010). Alternatively, the alignment of responses for Muslim and left-leaning persona might come from an epistemic overlap (e.g., postcolonial scholarship emerging from historically colonized Muslim-majority regions (Meer, 2014)) that the model reproduces. Regardless of interpretation, these findings underscore the importance of context-aware alignment: emancipatory approach to epistemic justice must be grounded in the sociopolitical realities of the community in question. Without such grounding, LLMs risk reproducing ideologies that are centered around justice in one setting but hegemonic in another. Therefore, an alignment framework should not assume universal moral or political priors, but instead incorporate historically and culturally situated knowledge–especially when engaging with the perspectives of marginalized and minority communities.

6 Conclusion

In examining how GPT-4's responses to Bengali political discourse deviate from its baseline responses while adopting different political and demographic personas, we found that it exhibits a measurable left-leaning bias. Although we did not find a significant relationship between the generated texts' political leanings and the topics or most demographic personas, only the majority Muslim identity-based persona produced responses that were significantly aligned with a political orientation. These tendencies carry major implications for how culture and society are (re)constructed through LLMs and generative AI. As these increasingly shape global cultural production, their alignment with dominant identities risks enforcing cultural and ideological homogeneity across languages and contexts, and contributing to the gradual disappearance of dissenting or minority views. These findings underscore the importance of auditing LLMs that take into account sociopolitical and cultural contexts in underrepresented languages and intersectionally diverse communities, thereby preventing the erasure of minority and marginalized perspectives. We call for greater attention to the epistemic impact of model alignment and for frameworks to evaluate political and identity biases in the Global South, non-Western, and

low-resource contexts.

7 Limitations

This paper offers insights into the sociopolitical alignment of LLM-generated texts in Bengali. However, in this section, we reflect on several limitations of our study. First, while some prior studies (Rozado, 2024; Thapa et al., 2023) advocate for incorporating both the left–right and authoritarian–libertarian dimensions to capture political orientation, our study focuses solely on the former–following precedent in much of the NLP literature (Li and Goldwasser, 2021; Motoki et al., 2024). As a result, it does not account for the additional ideological variation captured by the latter, which may be particularly relevant in the context of South Asian political discourse. Second, as we compare the similarities between the left- and right-leaning responses to the baseline response, our operationalization of political leaning becomes effectively binary. Moreover, we limit our analysis to two dominant religious (Hindu and Muslim) and national (Bangladeshi and Indian) identities within Bengali communities–such binarification overlooks the broader spectrum of political and cultural affiliations, particularly among smaller minority groups. Third, while we examine different religion and nationality categories separately, our study does not account for intersectional identities (e.g., Bangladeshi Muslims vs. Indian Hindus), which may exhibit distinct discursive patterns. Fourth, other key sociocultural dimensions such as gender, caste, and linguistic subregionality are not considered, despite their centrality in shaping Bengali political expression. Fifth, we used a multilingual model to generate the embeddings. While it performs better than models trained only using English data on Bengali texts, it generally underperforms compared to models pretrained exclusively on Bengali or other closely related languages (Das et al., 2025b; Ogunremi et al., 2023). As a result, the embeddings may suffer from contextual loss or reduced linguistic nuance. Sixth, the dataset we used primarily reflects discourse within the national contexts of Bangladesh and India, with less explicit attention to diasporic Bengali communities whose perspectives may differ due to transnational experiences. Finally, this paper audits the biases in GPT-4 by OpenAI. While it is one of the most widely used LLM (Chen et al., 2024), future work should examine biases in a wider array of LLMs and propose bias mitigation strategies in regards to the complexity and diversity of sociopolitical identities in Bengali discourse.

Ethical Considerations

In this section, we reflect on the ethical considerations, objectives, and scope of our study in light of a recent controversy in AI research and in relation to our own positionality as researchers.

Research Objective and Scope

Our study analyzed LLM responses to prompts combining lab-constructed personas with posts from BTPD (Das et al., 2025a). While the dataset includes content collected from online platforms, we did not post any generated responses back or engage with users in those communities. This stands in contrast to recent ethically controversial studies–such as the experiment involving undisclosed AI-generated responses on Reddit–which violated community norms and user trust by deploying persuasive bots in real time (IE et al., 2025). In our case, we conducted all analyses offline, and limited the use of community-sourced data to prompt design. We did not make any interventions in the platforms from which data was sourced, and did not make any attempts to deceive, persuade, or manipulate users. Additionally, we followed established ethical guidelines for research involving publicly available social media data (Fiesler and Proferes, 2018), including not using usernames and other sensitive or personally identifiable content. Our goal was to understand how LLMs reflect or prioritize sociopolitical perspectives in a controlled, non-interactive setting that preserves the integrity of the original online communities.

Positionality Statement

Researchers' identities may reflexively address inevitable tensions and bring affinities into perspective in studying underrepresented communities like the Bengalis (Schlesinger et al., 2017; Liang et al., 2021). Given this paper's focus on religion and nationality, we reflect here on the authors' identities across these dimensions. The first author was born and raised in Bangladesh in a Hindu family belonging to an underprivileged caste minority. The second author also grew up in Bangladesh, in a Muslim household. The third author was raised in India in a Hindu family. All authors (heterosexual men) are researchers at a North American university and have backgrounds in computer and information science, with prior research experi-

ence with marginalized communities and human-centered data science, which have informed and guided the motivation and execution of this study.

Acknowledgment

We used Grammarly Premium during our writing.

References

Utkarsh Agarwal, Kumar Tanmay, Aditi Khandelwal, and Monojit Choudhury. 2024. Ethical reasoning and moral value alignment of llms depend on the language we prompt them in. *arXiv preprint arXiv:2404.18460*.

Samyak Agrawal, Kshitij Gupta, Devansh Gautam, and Radhika Mamidi. 2022. Towards detecting political bias in Hindi news articles. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 239–244, Dublin, Ireland. Association for Computational Linguistics.

Alan Agresti. 2013. *Categorical data analysis*. John Wiley & Sons.

Senthil Kumar B, Pranav Tiwari, Aman Chandra Kumar, and Aravindan Chandrabose. 2022. Casteism in India, but not racism - a study of bias in word embeddings of Indian languages. In *Proceedings of the First Workshop on Language Technology and Resources for a Fair, Inclusive, and Safe Society within the 13th Language Resources and Evaluation Conference*, pages 1–7, Marseille, France. European Language Resources Association.

Ricardo Baeza-Yates. 2020. Bias in search and recommender systems. In *Proceedings of the 14th ACM Conference on Recommender Systems*, pages 2–2.

Victoria Balan and Delia Dumitrica. 2024. Technologies of last resort: The discursive construction of digital activism in wired and time magazine, 2010–2021. *new media & society*, 26(9):5466–5485.

Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. 2020. We can detect your bias: Predicting the political ideology of news articles. *arXiv preprint arXiv:2010.05338*.

Yejin Bang, Delong Chen, Nayeon Lee, and Pascale Fung. 2024. Measuring political bias in large language models: What is said and how it is said. *arXiv preprint arXiv:2403.18932*.

Marianne Bertrand and Sendhil Mullainathan. 2004. Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *American economic review*, 94(4):991–1013.

Shaily Bhatt, Sunipa Dev, Partha Talukdar, Shachi Dave, and Vinodkumar Prabhakaran. 2022. Recontextualizing fairness in NLP: The case of India. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 727–740, Online only. Association for Computational Linguistics.

Avigyan Bhattacharya, Tapabrata Chakrabarti, Subhadip Basu, Alistair Knott, Dino Pedreschi, Raja Chatila, Susan Leavy, David Eyers, Paul D Teal, and Przemyslaw Biecek. 2024. Towards a crowdsourced framework for online hate speech moderation-a case study in the indian political scenario. In *Companion Publication of the 16th ACM Web Science Conference*, pages 75–84.

Rishi Bommasani and Percy Liang. 2024. Trustworthy social bias measurement. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 210–224.

Kyle Booten. 2016. Hashtag drift: Tracing the evolving uses of political hashtags over time. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 2401–2405.

Andrei Boutyline and Robb Willer. 2017. The social structure of political echo chambers: Variation in ideological homophily in online networks. *Political psychology*, 38(3):551–569.

Bangladesh Statistics Bureau BSB. 2022. Preliminary report on population and housing census 2022 : English version. https://drive.google.com/file/d/1Vhn2t_PbEzo5-NDGBeoFJq4XCoSzOVKg/view. Last accessed: Feb 28, 2023.

Sébastien Bubeck, Varun Chadrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4.

Taina Bucher. 2012. Want to be on the top? algorithmic power and the threat of invisibility on facebook. *New media & society*, 14(7):1164–1180.

Maarten Buyl, Alexander Rogiers, Sander Noels, Guillaume Bied, Iris Dominguez-Catena, Edith Heiter, Iman Johary, Alexandru-Cristian Mara, Raphaël Romero, Jefrey Lijffijt, et al. 2024. Large language models reflect the ideology of their creators. *arXiv preprint arXiv:2410.18417*.

Bianca Chan and Reed Alexander. 2025. How goldman sachs is assembling a growing arsenal of ai tools: Here's everything we know about 5. *Business Insider*.

Partha Chatterjee. 2020. The nation and its fragments: Colonial and postcolonial histories.

Emily Chen, Ashok Deb, and Emilio Ferrara. 2022. # election2020: the first public twitter dataset on the 2020 us presidential election. *Journal of Computational Social Science*, pages 1–18.

Le Chen, Ruijun Ma, Anikó Hannák, and Christo Wilson. 2018. Investigating the impact of gender on rank in resume search engines. In *Proceedings of the 2018 chi conference on human factors in computing systems*, pages 1–14.

Lingjiao Chen, Matei Zaharia, and James Zou. 2024. How is chatgpt's behavior changing over time? *Harvard Data Science Review*, 6(2). Accessed: 2025-05-19.

John Cheney-Lippold. 2017. We are data. In *We Are Data*. New York University Press.

Matthew B Crawford. 2019. Algorithmic governance and political legitimacy. *American Affairs*, 3(2):73–94.

Dipto Das, Syed Ishtiaque Ahmed, and Shion Guha. 2025a. Btpd: A multilingual hand-curated dataset of bengali transnational political discourse across online communities. In *Companion Publication of the 2025 Conference on Computer-Supported Cooperative Work and Social Computing*, pages 188–193.

Dipto Das, Dhwani Gandhi, and Bryan Semaan. 2024a. Reimagining communities through transnational bengali decolonial discourse with youtube content creators. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW2):1–36.

Dipto Das, Shion Guha, Jed R Brubaker, and Bryan Semaan. 2024b. The"colonial impulse" of natural language processing: An audit of bengali sentiment analysis tools and their identity-based biases. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–18.

Dipto Das, Shion Guha, and Bryan Semaan. 2025b. How do datasets, developers, and models affect biases in a low-resourced language?

Dipto Das, AKM Najmul Islam, SM Taiabul Haque, Jukka Vuorinen, and Syed Ishtiaque Ahmed. 2022. Understanding the strategies and practices of facebook microcelebrities for engaging in sociopolitical discourses. In *Proceedings of the 2022 International Conference on Information and Communication Technologies and Development*, pages 1–19.

Dipto Das, Carsten Østerlund, and Bryan Semaan. 2021. " jol" or" pani"?: How does governance shape a platform's identity? *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–25.

Dipto Das and Bryan Semaan. 2022. Collaborative identity decolonization as reclaiming narrative agency: Identity work of bengali communities on quora. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–23.

Maryam Davoodi, Eric Waltenburg, and Dan Goldwasser. 2020. Understanding the language of political agreement and disagreement in legislative texts. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5358–5368.

Mark Díaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. 2018. Addressing age-related bias in sentiment analysis. In *Proceedings of the 2018 chi conference on human factors in computing systems*, pages 1–14.

Afia Dil. 1972. *The Hindu and Muslim Dialects of Bengali*. Stanford University.

Bryan Dosono and Bryan Semaan. 2018. Identity work as deliberation: Aapi political discourse in the 2016 us presidential election. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–12.

Paul Dourish and Scott D Mainwaring. 2012. Ubicomp's colonial impulse. In *Proceedings of the 2012 ACM conference on ubiquitous computing*, pages 133–142.

Benjamin G Edelman and Michael Luca. 2014. Digital discrimination: The case of airbnb. com. *Harvard Business School NOM Unit Working Paper*, (14-054).

The Editors of Encyclopædia Britannica. 2025. Bengali. Accessed: 2025-05-17.

Casey Fiesler and Nicholas Proferes. 2018. "participant" perceptions of twitter research ethics. *Social Media+ Society*, 4(1):2056305118763366.

Claudia Flores-Saviaga, Brian Keegan, and Saiph Savage. 2018. Mobilizing the trump train: Understanding collective action in a political trolling community. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Miranda Fricker. 2007. *Epistemic injustice: Power and the ethics of knowing*. Oxford University Press.

Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM Transactions on information systems (TOIS)*, 14(3):330–347.

Suyash Fulay, William Brannon, Shrestha Mohanty, Cassandra Overney, Elinor Poole-Dayan, Deb Roy, and Jad Kabbara. 2024. On the relationship between truth and political bias in language models. *arXiv preprint arXiv:2409.05283*.

Daniel Halpern and Jennifer Gibbs. 2013. Social media as a catalyst for online deliberation? exploring the affordances of facebook and youtube for political expression. *Computers in human behavior*, 29(3):1159–1168.

Brandon C Harris, Maxwell Foxman, and William C Partin. 2023. "don't make me ratio you again": How political influencers encourage platformed political participation. *Social Media+ Society*, 9(2):20563051231177944.

Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte. 2023. The political ideology of conversational ai: Converging evidence on chatgpt's pro-environmental, left-libertarian orientation. *arXiv preprint arXiv:2301.01768*.

Ellen Hoffman. 2024. Study: Some language reward models exhibit political bias. *MIT News*. Accessed: 2025-05-17.

David W Hosmer, Stanley Lemeshow, and Rodney X Sturdivant. 2013. *Applied logistic regression*. John Wiley & Sons.

Tenghao Huang, Faeze Brahman, Vered Shwartz, and Snigdha Chaturvedi. 2021. Uncovering implicit gender bias in narratives through commonsense inference. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3866–3873, Punta Cana, Dominican Republic. Association for Computational Linguistics.

W IE, G MA, and Y IMA. 2025. 'unethical'ai research on reddit under fire. *Science*.

Census India. 2011. Census tables. https://censusindia.gov.in/census.website/data/census-tables. Last accessed: Feb 28, 2023.

Lilly Irani, Janet Vertesi, Paul Dourish, Kavita Philip, and Rebecca E Grinter. 2010. Postcolonial computing: a lens on design and development. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1311–1320.

Ian T Jolliffe. 2002. *Principal component analysis for special types of data*. Springer.

Prerna Juneja and Tanushree Mitra. 2021. Auditing e-commerce platforms for algorithmically curated vaccine misinformation. In *Proceedings of the 2021 chi conference on human factors in computing systems*, pages 1–27.

Haesung Jung, Wenhao Dai, and Dolores Albarracín. 2024. How social media algorithms shape offline civic participation: A framework of social-psychological processes. *Perspectives on Psychological Science*, 19(5):767–780.

Will Knight. 2024. Chatbot teamwork makes the ai dream work. *WIRED*.

Sachin Kumar, Shuly Wintner, Noah A Smith, and Yulia Tsvetkov. 2019. Topics to avoid: Demoting latent confounds in text classification. *arXiv preprint arXiv:1909.00453*.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. *Advances in neural information processing systems*, 30.

George Lakoff. 2016. *Moral politics: How liberals and conservatives think*. University of Chicago Press.

Daniel D Lee and H Sebastian Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *nature*, 401(6755):788–791.

Chang Li and Dan Goldwasser. 2021. Using social and linguistic information to adapt pretrained representations for political perspective identification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4569–4579.

Calvin A Liang, Sean A Munson, and Julie A Kientz. 2021. Embracing four tensions in human-computer interaction research with marginalized people. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 28(2):1–47.

Ruibo Liu, Chenyan Jia, Jason Wei, Guangxuan Xu, and Soroush Vosoughi. 2022. Quantifying and alleviating political bias in language models. *Artificial Intelligence*, 304:103654.

Leland McInnes, John Healy, and James Melville. 2018. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Nasar Meer. 2014. Islamophobia and postcolonialism: continuity, orientalism and muslim consciousness. *Patterns of Prejudice*, 48(5):500–515.

Danaë Metaxa, Joon Sung Park, Ronald E Robertson, Karrie Karahalios, Christo Wilson, Jeff Hancock, Christian Sandvig, et al. 2021. Auditing algorithms: Understanding algorithmic systems from the outside in. *Foundations and Trends® in Human–Computer Interaction*, 14(4):272–344.

Abir Mondal, Kingshuk Roy, Susmita Das, and Arpita Dutta. 2024. Detecting toxic comments in bengali language. In *International Conference on Computational Intelligence in Pattern Recognition*, pages 557–568. Springer.

Fabio Motoki, Valdemar Pinho Neto, and Victor Rodrigues. 2024. More human than human: measuring chatgpt political bias. *Public Choice*, 198(1):3–23.

Andrew Ng, Rohit Prasad, Kevin Solorio, Ryan Prinz, Jeff Tang, Riddhimaan Senapati, Christopher Michael-Stokes, John Santerre, Kamilk Cerebras, Zoltan Csaki, Rohit, Dax Patel, Evan d'Entremont, Ming Gong, Yuan Man, Gautam Goudar, Bilal Hamada, Ikko Eltociear Ashimine, Hatice Ozen, Aditya Rana, Lucain, Adarsh Shirawalmath, Kevin Bazira, Neel Patel, BRlin-o, and Isaac Tian. 2024. AISuite: A Modular Framework for AI Workflows. https://github.com/andrewyng/aisuite. Accessed: 2025-04-17.

Tolulope Ogunremi, Dan Jurafsky, and Christopher D Manning. 2023. Mini but mighty: Efficient multilingual pretraining with linguistically-informed data selection. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1251–1266.

Hashai Papneja and Nikhil Yadav. 2024. Self-disclosure to conversational ai: A literature review, emergent framework, and directions for future research. *Personal and ubiquitous computing*, pages 1–33.

Weihong Qi, Hanjia Lyu, and Jiebo Luo. 2024. Representation bias in political sample simulations with large language models. *arXiv preprint arXiv:2407.11409*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Luca Rettenberger, Markus Reischl, and Mark Schutera. 2025. Assessing political bias in large language models. *Journal of Computational Social Science*, 8(2):1–17.

Rodrigue Rizk, Dominick Rizk, Frederic Rizk, and Sonya Hsu. 2023. 280 characters to the white house: predicting 2020 us presidential elections from twitter data. *Computational and Mathematical Organization Theory*, 29(4):542–569.

Ronald E Robertson, Shan Jiang, Kenneth Joseph, Lisa Friedland, David Lazer, and Christo Wilson. 2018. Auditing partisan audience bias within google search. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW).

David Rozado. 2024. The political preferences of llms. *PloS one*, 19(7):e0306621.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1668–1678.

Morgan Klaus Scheuerman, Alex Hanna, and Emily Denton. 2021. Do datasets have politics? disciplinary values in computer vision dataset development. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–37.

Morgan Klaus Scheuerman, Jacob M Paul, and Jed R Brubaker. 2019. How computers see gender: An evaluation of gender classification in commercial facial analysis services. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–33.

Ari Schlesinger, W Keith Edwards, and Rebecca E Grinter. 2017. Intersectional hci: Engaging identity through gender, race, and class. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, pages 5412–5427.

Bryan C Semaan, Scott P Robertson, Sara Douglas, and Misa Maruyama. 2014. Social media supporting political deliberation across multiple public spheres: towards depolarization. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 1409–1421.

Samuel Sanford Shapiro and Martin B Wilk. 1965. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4):591–611.

SIL International. 2023. Ethnologue 200: The top 200 most spoken languages. Accessed: 2025-05-17.

Kate Starbird and Leysia Palen. 2012. (how) will the revolution be retweeted? information diffusion and the 2011 egyptian uprising. In *Proceedings of the acm 2012 conference on computer supported cooperative work*, pages 7–16.

Student. 1908. The probable error of a mean. *Biometrika*, pages 1–25.

Henri Tajfel. 1974. Social identity and intergroup behaviour. *Social science information*, 13(2):65–93.

Nazia Tasnim, Sujan Sen Gupta, Md Istiak Hossain Shihab, Fatiha Islam Juee, Arunima Tahsin, Pritom Ghum, Kanij Fatema, Marshia Haque, Wasema Farzana, Prionti Nasir, et al. 2024. Mapping violence: Developing an extensive framework to build a bangla sectarian expression dataset from social media interactions. *arXiv preprint arXiv:2404.11752*.

Zerin Tasnim, Shuvo Ahmed, Atikur Rahman, Jannatul Ferdous Sorna, and Mafizur Rahman. 2021. Political ideology prediction from bengali text using word embedding models. In *2021 international conference on emerging smart computing and informatics (ESCI)*, pages 724–727. IEEE.

Surendrabikram Thapa, Ashwarya Maratha, Khan Md Hasib, Mehwish Nasim, and Usman Naseem. 2023. Assessing political inclination of bangla language models. In *BLP 2023-1st Workshop on Bangla Language Processing, Proceedings of the Workshop*, pages 152–162. Association for Computational Linguistics.

Samia Touileb, Lilja Øvrelid, and Erik Velldal. 2022. Occupational biases in Norwegian and multilingual language models. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 200–211, Seattle, Washington. Association for Computational Linguistics.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Pranav Narayanan Venkit, Sanjana Gautam, Ruchi Panchanadikar, Shomir Wilson, et al. 2023. Nationality bias in text generation. *arXiv preprint arXiv:2302.02463*.

Pranav Narayanan Venkit, Mukund Srinath, and Shomir Wilson. 2022. A study of implicit bias in pretrained language models against people with disabilities. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1324–1332, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Frank Wilcoxon. 1992. Individual comparisons by ranking methods. In *Breakthroughs in statistics: Methodology and distribution*, pages 196–202. Springer.

Denise J Wilkins, Andrew G Livingstone, and Mark Levine. 2019. Whose tweets? the rhetorical functions of social media use in developing the black lives matter movement. *British Journal of Social Psychology*, 58(4):786–805.

Alex J Wood, Mark Graham, Vili Lehdonvirta, and Isis Hjorth. 2019. Good gig, bad gig: autonomy and algorithmic control in the global gig economy. *Work, employment and society*, 33(1):56–75.

Dogus Yuksel, Mehmet Cem Catalbas, and Bora Oc. 2025. Language-dependent political bias in ai: A study of chatgpt and gemini. *arXiv preprint arXiv:2504.06436*.