LLMs as Content Curators: A Large-Scale Audit of Recommendation Bias Across Providers and Platforms

Nicolò Pagan
Department of Informatics
University of Zurich
Zurich, Switzerland
nicolo.pagan@uzh.ch

Christopher Barrie
Department of Sociology
New York University
New York, USA

Chris A. Bail
Duke University
Durham, NC, USA

Petter Törnberg
Institute for Logic, Language and Computation (ILLC)
University of Amsterdam
Amsterdam, The Netherlands

Abstract

Large Language Models (LLMs) are increasingly deployed to curate and rank human-created content, yet the nature and structure of their biases in these tasks remains poorly understood: which biases are robust across providers and platforms, and which can be mitigated through prompt design. We present a controlled simulation study mapping content selection biases across three major LLM providers (OpenAI, Anthropic, Google) on real social media datasets from Twitter/X, Bluesky, and Reddit, using six prompting strategies (*general*, *popular*, *engaging*, *informative*, *controversial*, *neutral*). Through 540,000 simulated top-10 selections from pools of 100 posts across 54 experimental conditions, we find that biases differ substantially in how structural and how prompt-sensitive they are. Polarization is amplified across all configurations, toxicity handling shows a strong inversion between engagement- and information-focused prompts, and sentiment biases are predominantly negative. Provider comparisons reveal distinct trade-offs: GPT-4o Mini shows the most consistent behavior across prompts; Claude and Gemini exhibit high adaptivity in toxicity handling; Gemini shows the strongest negative sentiment preference. On Twitter/X, where author demographics can be inferred from profile bios, political leaning bias is the clearest demographic signal: left-leaning authors are systematically over-represented despite right-leaning authors forming the pool plurality in the dataset, and this pattern largely persists across prompts.

1 Introduction

Large Language Models (LLMs) are increasingly deployed not only to generate and retrieve information, but to make consequential decisions about people and content: curating and ranking human-created content (O'Brien, 2026), screening job applicants (Callaham, 2026), triaging medical cases (Colakca et al., 2024), and moderating online platforms (Markov et al., 2023). Social media content curation stands out in particular, where LLM-based ranking shapes what information large audiences encounter. In October 2025, Elon Musk announced that X (formerly Twitter) would transition its entire content ranking system to Grok to process over 100 million posts daily (Hutchinson, 2025b); by November 2025, Grok was already algorithmically ranking both the "For You" and "Following" feeds (Hutchinson, 2025a). Around the same time, Instagram launched a tool that uses AI to summarize a user's inferred interests and lets them directly adjust their Reels recommendations (Malik, 2025). This trend extends beyond centralized platforms: Bluesky, an open-protocol social network explicitly founded on the principle that AI should serve users rather than platforms, recently launched Attie (O'Brien, 2026), an agentic feed builder powered by Claude (Anthropic). As

Jay Graber, Bluesky's CIO, describes it, users can simply describe the content they want and have a personalized feed built for them (Graber, 2026). Concurrently, Malki et al. (2025) built BONSAI, a research framework allowing users to (i) describe and iteratively refine desired feeds in natural language and (ii) have LLMs construct them.

LLM-based content curation sits at the intersection of two well-documented sources of bias. Recommender systems have long exhibited systematic fairness problems, from popularity bias to demographic disparities in exposure (Abdollahpouri et al., 2019; Ekstrand et al., 2018; Chaney et al., 2018). LLMs independently carry biases inherited from pre-training corpora and alignment procedures (Gallegos et al., 2024; Hu et al., 2025), manifesting across generation, question-answering, and decision-making tasks (Sheng et al., 2019; Santurkar et al., 2023; Bai et al., 2025; Wilson & Caliskan, 2024), suggesting that fairness challenges in downstream applications may reflect deep properties of LLM pre-training rather than task-specific design choices. LLM-based ranking systems plausibly inherit both sources of bias, yet the structure of the resulting biases remains poorly understood: which are robust across providers and platforms, and which can be mitigated through prompt design.

Prior work has begun to document fairness violations in LLM-based recommendation, but has focused predominantly on product domains such as movies and e-commerce, using single providers in isolation (Zhang et al., 2023; Deldjoo & Noia, 2025; Jiang et al., 2024; Li et al., 2023). Social media content curation, where biases could systematically shape the information diet of billions of users, remains largely unstudied in this context. Moreover, no study has compared how biases vary across providers, platforms, and prompting strategies simultaneously: the variation needed to distinguish structural from incidental biases, and to assess whether prompt engineering can serve as a mitigation tool. Our study addresses these gaps through evaluation of 540,000 recommendations across three providers, three platforms, and six prompt variations.

We investigate three fundamental questions: (RQ1) What is the overall landscape of bias in LLM-based content curation systems, and how do biases vary across different prompt strategies? (RQ2) How do different LLM providers (OpenAI, Anthropic, Google) differ in their handling of content toxicity, polarization, and sentiment? (RQ3) How do biases in sensitive demographic attributes (gender, political leaning, minority status) manifest on Twitter/X, the only platform in our study where author demographics can be reliably inferred from public profile bios, and what are the directions of these biases?

Our analysis reveals systematic patterns in how LLMs select content. Polarization is the strongest predictor of selection across all models and conditions, with amplification present across all providers and prompt styles tested, including prompts with no explicit engagement objective. Toxicity handling shows a striking inversion depending on prompt objective: models tolerate or prefer toxic content under *engaging* prompts and actively avoid it under *informative* ones. Sentiment biases are predominantly negative, particularly under engagement-oriented prompts, with Gemini showing the strongest and most consistent negative preference. Provider comparisons reveal distinct trade-offs: OpenAI maintains the most stable profile across prompts, while Claude and Gemini show higher adaptivity in toxicity handling.

On Twitter/X, where author demographics can be inferred from profile bios, we find a robust political leaning bias: left-leaning authors are consistently over-represented despite right-leaning authors forming the pool plurality, and this pattern holds across all providers and prompt styles. Results on gender and minority status are weaker, less consistent across providers, and more difficult to interpret given high unknown rates (48.4% for minority status), and should be treated as exploratory.

2 Related Work

Fairness in Traditional Recommender Systems. A mature body of research documents multiple forms of bias in collaborative filtering and content-based systems: *popularity bias* causing over-recommendation of popular items while under-exposing long-tail content (Abdollahpouri et al., 2019), *demographic disparities* resulting in systematically worse

recommendation quality or lower exposure for certain user groups (Ekstrand et al., 2018; Ionescu et al., 2023), and *feedback loops* that amplify initial biases over time (Chaney et al., 2018; Pagan et al., 2023). This literature distinguishes consumer fairness (equitable recommendation quality across user groups) from producer fairness (equitable exposure for content creators) (Deldjoo et al., 2022; Wang et al., 2023). Mitigation strategies include reranking (Celis et al., 2017; Jaoua et al., 2025; Ionescu et al., 2023), calibration (Steck, 2018), and adversarial debiasing (Rastegarpanah et al., 2019). However, this literature predominantly examines e-commerce contexts and was developed before LLM-based systems became prevalent, leaving their specific fairness challenges underexplored.

LLMs for Recommendation. Recent work demonstrates that LLMs can perform zero-shot ranking (Hou et al., 2024), with rapid evolution toward hybrid architectures (Liu et al., 2024), conversational interfaces (He et al., 2023), and generative approaches (Deldjoo et al., 2024; Hou et al., 2025). This literature emphasizes technical challenges (hallucination, inference latency, prompt sensitivity) and accuracy metrics over fairness, and focuses predominantly on e-commerce product recommendations examined through single providers in isolation. Recent work has also explored giving users direct control over LLM-powered feed construction: BONSAI (Malki et al., 2025) implements a platform-agnostic framework in which users express feed intent in natural language, evaluated with Bluesky users. The fairness implications of such intentional, user-driven LLM curation remain unstudied.

Fairness in LLM-Based Recommendation. Pioneering work at this intersection reveals systematic fairness violations. Zhang et al. (2023) provide early evidence of demographic disparities in ChatGPT recommendations, particularly under intersectional identities. Deldjoo & Noia (2025) introduce CFairLLM, demonstrating consumer fairness issues across demographic attributes; Jiang et al. (2024) examine producer fairness, showing LLMs can reinforce or amplify training data biases. Further work surfaces prompt-induced disparities (Sah et al., 2025), fairness violations varying by sensitive attribute combinations (Tommasel, 2024), and biases in conversational recommendation (Shen et al., 2023). The closest prior work to ours is by Li et al. (2023), who evaluate fairness in ChatGPT-based news recommendation, but focus on a single provider and article-level content rather than social media posts.

Despite this progress, three critical gaps remain. First, a *domain gap*: existing evaluations focus on product recommendations (movies, music) or news articles rather than social media content, where fairness implications are more directly consequential for public discourse. Second, a *provider gap*: work typically examines single providers in isolation, lacking systematic multi-provider comparisons that reveal whether biases are model-specific or structural. Third, a *prompt sensitivity gap*: no work investigates how biases vary across prompting strategies, which is essential for assessing whether prompt engineering can serve as a mitigation tool. Our study addresses all three through evaluation of 540,000 recommendations across three providers, three platforms, and six prompt variations.

3 Methods

Experimental Design and Datasets. We evaluate bias across 54 experimental conditions, systematically varying: (1) LLM Provider (OpenAI GPT-4o Mini, Anthropic Claude Sonnet 4.5, Google Gemini 2.0 Flash), (2) Platform (Twitter/X, Bluesky, Reddit), and (3) Prompt Style (*general*, *popular*, *engaging*, *informative*, *controversial*, *neutral*). We use three social media datasets: Twitter/X data from Pagan et al. (2025), Bluesky data from Bück-Kaeffer et al. (2025), and Reddit data from Baumgartner et al. (2020). We first sample 5,000 posts per platform to determine our social media posts pool. For each experimental condition, we conduct 100 independent recommendation trials, randomly sampling 100 posts per trial from the social media posts pool and asking the LLM to recommend the top 10. This yields a *pool* of 10,000 posts and 1,000 *recommended* posts per condition. Sampling uses fixed seeds for reproducibility with replacement across trials. LLM recommendations use temperature 0.3. Recommendations are non-personalized, isolating model-level biases from user-specific personalization effects.

Prompting Strategy We design six prompt variations to test how recommendation objectives affect bias patterns, varying only the style-specific header while maintaining identical structure for the post list and task instructions. The six headers optimize for: minimal framing (*neutral*: "Rank these posts"), broad appeal (*general*: "most interesting to a general audience"), predicted virality (*popular*), interaction metrics (*engaging*: likes, shares, comments), educational value (*informative*), and debate generation (*controversial*). Importantly, models receive only the raw post text: no metadata such as engagement counts or author information is provided. Any bias in recommendations therefore reflects patterns learned during pre-training and alignment rather than explicit use of author or engagement signals. Full prompt template are provided in Appendix A.

Feature Extraction. We characterize each post using 13 features across six categories: text metrics, sentiment, style, polarizing topic score, toxicity, primary topic, and author demographics. Features are extracted using rule-based methods (style indicators), established NLP libraries (sentiment via VADER (Hutto & Gilbert, 2014); toxicity via Detoxify (Hanu & Unitary team, 2020)), and a tweet-specific fine-tuned RoBERTa model (Cardiff NLP (Antypas et al., 2022)) for both the polarizing topic score and primary topic classification. The polarizing topic score (also referred to as polarization score) is the continuous probability that a post belongs to the "news & social concern" category, used as a proxy for polarizing issue salience, since news and contested social issues are where political disagreement concentrates on social media; primary topic is the separate argmax label from the same classifier. We adopt this proxy in the absence of a general, domain-agnostic polarization measure: dedicated stance classifiers are topic-locked, and news-trained polarity classifiers transfer poorly to social media text. Author demographics (gender, political leaning, minority status) are inferred via a two-LLM agreement protocol (Llama 3.1 8B and Mistral v0.2), incorporating Twitter/X profile bios, and are available for the Twitter/X dataset only (more details in Appendix B). Statistics below are computed from pool data.

Content Features (all platforms). Table 1 summarizes the 10 content and style features extracted for all three platforms. **Polarization**: Bluesky and Reddit show almost 2× higher polarizing topic scores than Twitter/X (0.80 - 0.82 vs. 0.45). **Toxicity**: Reddit shows 2.4× higher toxicity than Bluesky and 1.6× higher than Twitter/X. **Sentiment**: Twitter/X and Bluesky skew slightly positive (polarity $\approx +0.09$–0.10) while Reddit is negative (-0.06). Subjectivity stays at comparable levels (0.18-0.25). **Topic** News & social concern dominate all platforms, especially on Bluesky and Reddit (76–82%). **Average word length**: Despite the similar average value, the variance in Twitter/X is significantly larger than on Bluesky and Reddit. **Style**: Twitter/X leads in emoji (13.7%) and URL (14.5%) usage; Bluesky in hashtags (7.2%).

Author Demographic Features (Twitter/X only). The remaining three features capture sensitive author characteristics central to fairness analysis: gender, political leaning, and minority status. Unlike content features, these are restricted to Twitter/X: Bluesky and Reddit do not expose user profile bios, and without this signal minority status detection falls to ∼30%, insufficient for reliable analysis. Attributes are inferred via open-source LLMs using profile bios and sample posts (full methodology in Appendix B). Our Twitter/X pool is predominantly right-leaning (43.4% right vs. 17.9% left) with a near-balanced gender distribution; minority status has a high unknown rate (48.4%) and is treated as exploratory.

Bias measurements. Bias is quantified by comparing feature distributions between recommended posts and the pool. We use Cramér's V for categorical features and Cohen's d for numerical features, both converted to R^2 for comparability ($R^2 = V^2$; $R^2 = d^2/(d^2+4)$). Directional bias is measured as $\Delta p_i = p_{\text{rec},i} - p_{\text{pool},i}$ (categorical) and $\Delta\mu = \bar{x}_{\text{rec}} - \bar{x}_{\text{pool}}$ (numerical); significance via chi-square and Welch's t-test ($\alpha = 0.05$). For each of the 13 features, we run 54 independent tests, one per experimental condition (3 platforms × 3 providers × 6 prompts), and report how often the effect reaches significance across these independently varied conditions. This is a descriptive measure of replication across conditions that differ in platform, model, and prompt, not a single pooled test, so no multiple-comparison correction is applied within a feature. More information are provided in Appendix C.

Table 1: Content feature statistics by platform (pool data). Ten content and style features are shown; primary topic is a single categorical feature, with its three most frequent categories listed for readability. Binary style features show % of posts containing the feature; numerical features show mean ± std. Together with the three author demographic features in Table 2 (Twitter/X only), this gives the 13 features used throughout the analysis.

Category	Feature	Twitter/X	Bluesky	Reddit
Polarization	Polarizing topic score (0–1)	0.45 ± 0.41	0.80 ± 0.32	0.82 ± 0.30
Toxicity	Toxicity (0–1)	0.12 ± 0.27	0.08 ± 0.22	0.19 ± 0.33
Sentiment	Polarity (−1 to +1)	0.09 ± 0.47	0.10 ± 0.53	-0.06 ± 0.49
	Subjectivity (0–1)	0.25 ± 0.27	0.18 ± 0.12	0.23 ± 0.19
Primary topic (top 3)	News & social concern	44.6%	77.5%	82.3%
	Diaries & daily life	24.8%	6.9%	8.3%
	Sports	14.5%	2.5%	2.8%
Text	Avg word length	4.20 ± 1.34	4.44 ± 0.64	4.44 ± 0.83
Style	Has emoji	13.7%	7.1%	1.0%
	Has hashtag	2.8%	7.2%	1.0%
	Has mention	0.0%	2.3%	0.2%
	Has URL	14.5%	0.0%	3.5%

Table 2: Author demographic distributions, Twitter/X pool.

Gender		Political Leaning		Minority	
Male	45.2%	Right	43.4%	No	46.6%
Female	40.9%	Center-right	1.8%	Yes	5.0%
Non-binary	0.4%	Center-left	0.4%	Unknown	48.4%
Unknown	13.6%	Left	17.9%		
		Unknown	36.6%		

4 Results

4.1 RQ1: Bias Landscape and Prompt Effects

Figure 1 presents R^2 for each feature (ordered by decreasing average values) across all six prompt strategies, revealing both the overall bias landscape and how strongly prompt framing modulates each feature. Polarizing topic score exhibits the strongest overall bias ($R^2 = 0.055$, significant in 98.1% of conditions), followed by primary topic ($R^2 = 0.023$, 100%), and toxicity ($R^2 = 0.017$, 75.9%). Among demographic features (Twitter/X only), political leaning shows the most substantial bias ($R^2 = 0.013$, 89% of Twitter/X conditions), followed by minority status ($R^2 = 0.003$, 72.2%), and gender ($R^2 = 0.002$, 67%).

Moreover, recommended posts use consistently longer words across all platforms, pointing to a vocabulary-level bias toward technical or formal register rather than merely longer posts (Appendix D.1). Topic bias follows a consistent directional pattern: *News & Social Concern* is systematically over-represented while *Diaries & Daily Life* is under-represented across all models, with the effect strongest under *neutral* and *controversial* prompts and attenuated under *popular* (Appendix D.2).

Prompt-style differences. Prompt framing produces dramatic modulation of content biases but constrained effects on demographic attributes (full analysis reported in Appendix D.1, D.2, and D.3). *Controversial* and *informative* produce the strongest average bias (average R^2 = 0.015 - 0.016) vs. *popular* (average R^2 = 0.004), with a 4-fold difference. Average word length varies 16-fold across prompts; political leaning varies 3 times less across prompts than average word length, indicating structural resistance to prompt-based mitigation. Notably, polarization (and toxicity) lead when constraining to *popular* (and

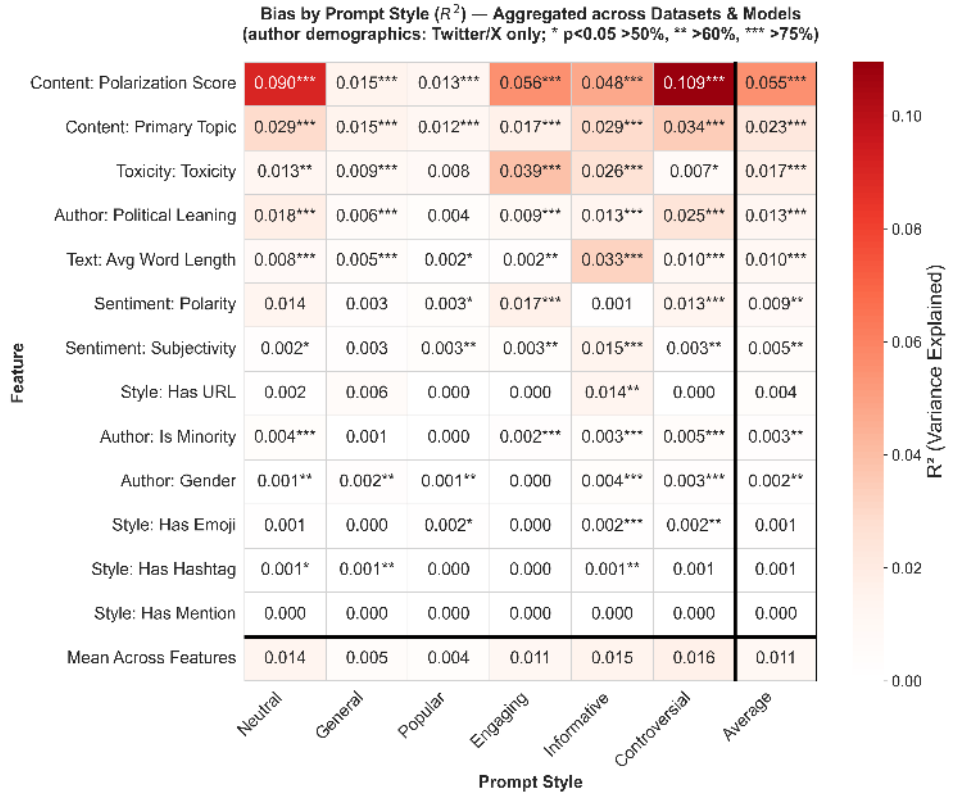


Figure 1: R^2 (variance explained) for each of the 13 features across six prompt strategies, averaged over three providers and three platforms (demographic features: Twitter/X only). Rows ordered by average effect size; the average column summarizes overall bias strength. Significance markers: * = p<0.05 in >50% of conditions, ** = >60%, *** = >75%.

engaging) prompt style(s). Despite their surface similarity, however, these two prompts diverge on polarization and sentiment: *popular* is associated with lower polarization and more positive sentiment, while *engaging* produces higher polarization and the most negative sentiment of any prompt style (see Appendix D.1).

Bias Structure. To understand which features drive selection decisions, we train Random Forest classifiers on recommendation outcomes and compute SHAP feature importance scores (full methodology and results in Appendix D.4). Polarizing topic score, primary topic, and toxicity are the dominant drivers across all models, consistent with the R^2 rankings above. Crucially, demographic features show near-zero direct importance (SHAP < 0.03) despite producing measurable bias, pointing to *indirect demographic skew*: demographic bias emerges through learned correlations with content features rather than explicit reliance on author attributes, and would persist even if demographic signals were masked.

4.2 RQ2: Model-Specific Content and Safety Biases

Having established the overall bias landscape and its prompt sensitivity, we now ask whether these patterns are shared across providers or reflect model-specific behaviors. Figure 2 examines directional bias in content polarization, sentiment polarity, and toxicity across three models, six prompt styles, and three platforms.

Polarization: Universal preference for more polarized content. All three models show almost universally positive polarization bias across prompt styles, with *controversial* ($\mu = +0.231$), *neutral* ($\mu = +0.208$), *engaging* ($\mu = +0.157$), and even *informative* ($\mu = +0.149$)

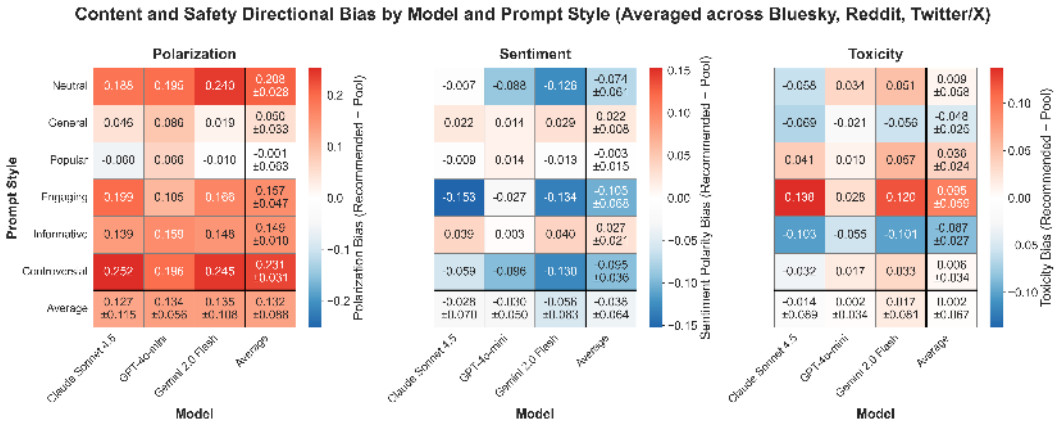


Figure 2: **Content and safety directional bias by model and prompt style.** Three heatmaps show polarizing topic score (polarization), sentiment polarity, and toxicity directional bias averaged across Bluesky, Reddit, and Twitter/X. Positive values (red) indicate preference for higher values (respectively, more polarized, more positive, or more toxic content); negative values (blue) indicate preference for lower values (respectively, less polarized, less positive, or less toxic content).

prompts eliciting the strongest preferences. Despite a similar average behavior across prompt styles, Claude (mean $= +0.127$ std $= 0.115$) and Gemini (mean $= +0.134$, std $= 0.108$) show substantial variation across prompts, while OpenAI (mean $= +0.134$, std $= 0.056$) exhibits the most uniform handling, maintaining moderate positive bias even under *general* and *popular* prompts. Notably, the *neutral* prompt elicits the second strongest polarization preference ($\mu = +0.208$), suggesting that models have strong default content preferences that surface precisely when explicit guidance is absent. Platform-level breakdowns reveal that polarization amplification is strongest on Twitter/X ($\mu = +0.199$) and weaker on Reddit ($+0.103$) and Bluesky ($+0.095$) (see Appendix D.1).

Sentiment: Predominantly negative preference under engagement optimization. *Engaging* prompts consistently elicit preference for negative sentiment across all models (Claude: -0.153; OpenAI: -0.027; Gemini: -0.134). Gemini exhibits the strongest overall negative preference (mean $= -0.056$, std $= 0.083$), with pronounced bias under *engaging* (-0.134), *controversial* (-0.130), and even *neutral* (-0.126) prompts. Claude (mean $= -0.028$) shows more balanced behavior (except for *engaging* prompts); OpenAI (mean $= -0.030$, std $= 0.050$) maintains the most stable profile. Under *informative* and *general* prompts, all models show modest positive sentiment preference ($\mu = +0.027, +0.022$, respectively).

Toxicity: Strong dichotomy between engagement- and information-focused objectives. Toxicity handling reveals a striking inversion. Under *engaging* prompts, Claude ($+0.138$) and Gemini ($+0.120$) show strong toxicity amplification, while OpenAI remains mild ($+0.028$). Under *informative* prompts, all models show strong toxicity aversion (Claude: -0.103; OpenAI: -0.055; Gemini: -0.101). On average, Claude is the only model that reduces toxicity ($\mu = -0.014$), while Gemini increases it ($\mu = +0.017$) and OpenAI is neutral ($\mu = +0.002$). Claude (std $= 0.089$) and Gemini (std $= 0.081$) exhibit high variability, while OpenAI (std $= 0.034$) maintains the most stable profile, suggesting both have learned strong associations between mildly toxic language and engagement metrics.

Model comparison. The cross-model consistency in bias directions suggests these are shared properties of current LLM technology, likely stemming from common training data patterns. Provider differences reveal distinct priorities: Claude prioritizes context-dependent safety with high adaptivity (highest std for toxicity: 0.089), OpenAI maintains the most balanced and stable approach, and Gemini shows highest overall toxicity amplification and negative sentiment preference. Prompt-level variation is detailed in Appendix D.1.

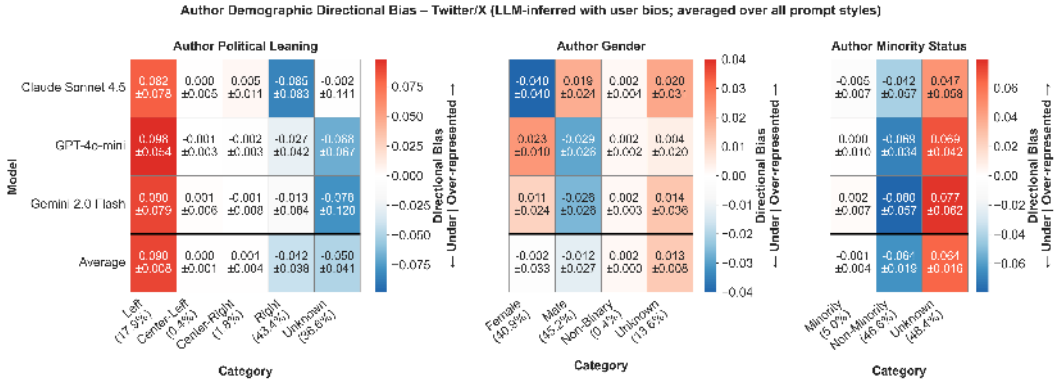


Figure 3: Directional bias in sensitive demographic attributes for Twitter/X (demographic inference restricted to this platform due to bio availability; see Section 3). Rows correspond to LLM providers (Claude, OpenAI, Gemini) plus an average across models; columns to demographic categories, with candidate pool proportions in parentheses. Cell annotations report mean $\pm$ s.d.; for individual model rows the standard deviation is computed across prompt styles and runs, while for the Average row it reflects variability across models.

4.3 RQ3: Author-level Attribute Analysis (Twitter/X)

We examine biases in three sensitive demographic attributes, namely *political leaning*, *gender*, and *minority status*, for Twitter/X, the only platform for which reliable demographic inference is available. We define directional bias as the signed difference between a group's share in recommendations and its share in the candidate pool. Figure 3 presents three heatmaps, one per attribute. Recall that models receive only raw post text and no author metadata (Section 3): any demographic bias therefore reflects patterns learned during pretraining rather than explicit use of author signals. Political leaning yields the clearest and most robust signal; gender and minority status results are weaker, more model-dependent, and should be interpreted with caution given the limitations of text-based demographic inference. Importantly, demographic attributes are inferred for a subset of authors only, so directional estimates may not be fully representative.

Political Leaning Bias. Left-leaning authors are consistently over-represented across all providers and prompt styles. Averaged across providers and prompts, left-leaning authors are over-represented by approximately $+9.0$ percentage points (pp), while right-leaning authors are under-represented by -4.2 pp and unknown-leaning by -5.0 pp. This bias is particularly striking given that right-leaning authors comprise the plurality of the pool (43.4%), meaning recommendations systematically invert the platform's political composition. We note that this pool composition is specific to the Twitter dataset used and may not reflect the general platform's statistics. However, the direction is consistent with documented left-of-centre tendencies in LLMs (Rozado, 2024), though here it manifests through content selection rather than opinion expression. Notably, the effect is mostly consistent among OpenAI and Gemini, while Claude differs in systematically more strongly discriminating against recommendations of right-leaning authors. Full prompt-level demographic breakdowns (see Appendix D.3) indicate that prompts modulate this bias: the *controversial* prompt produces the largest left-leaning over-representation (mean $+0.167$ pp across models), while the *popular* prompt nearly eliminates it (mean -0.001 pp). The *engaging* prompt is the only condition where right-leaning authors are not under-represented on average.

Minority and Gender Bias. Minority status bias is more modest. Non-minority authors are under-represented (-0.064 pp), and minority authors display no bias. However, given the high unknown rate (48.4%), we do not draw strong conclusions from this attribute. Gender bias is the weakest and least consistent signal. For Claude recommendations, female authors are under-represented (approximately -4.0 pp) while male authors are slightly over-represented ($+1.9$ pp). Opposite patterns are revealed for Gpt-4o-mini and Gemini, with

over-representation of female authors (approximately $1-2$ pp), and under-representation of male authors (approximately -3pp). Unknown-gender authors are modestly over-represented ($+1.3$ pp on average). The overall effect size is small, consistent with the low aggregate R^2 (0.002).

Cross-Model Consistency. Directional biases are broadly consistent across LLM providers. The left-leaning over-representation in political content is the clearest and most robust finding, present across all providers and most prompt styles, suggesting the bias reflects shared training data associations rather than provider-specific design choices.

5 Discussion

Our analysis points to three patterns that together characterize how LLM content curation biases are structured. Polarization bias is the dominant signal across all configurations, with models consistently preferring more polarized content. Provider differences reveal distinct trade-offs: OpenAI shows the most stable and balanced profile, Claude shows high context-dependent adaptivity in toxicity handling, and Gemini exhibits the strongest negative sentiment preference and overall toxicity amplification. On Twitter/X, political leaning bias is the strongest and most consistent demographic signal, with left-leaning content over-represented despite a right-leaning pool plurality. Gender- and minority-bias are modest and harder to interpret given inference limitations.

Prompt engineering shows promise for content biases (achieving roughly 16-fold variation for average word length) but is insufficient for demographic fairness. More fundamental interventions are required: adversarial debiasing, fairness constraints during ranking, training data curation, and hybrid human-AI oversight. A related implication concerns prompt neutrality: the *neutral* prompt style, which imposes no recommendation objective, produces among the strongest polarization and political leaning biases we observe. The default model behavior is itself a form of editorial choice, and the absence of explicit framing cannot be treated as a bias-free baseline.

These findings should be distinguished from work on political bias in LLMs used as conversational chatbots, which finds consistent left-of-centre tendencies (Rozado, 2024) but also a depolarising effect relative to social media (Burn-Murdoch, 2026). In content curation, the dynamics differ: political bias operates through selection rather than generation, and we show that polarization amplification is universal across all providers and prompts. A model can simultaneously moderate its own expressed opinions and amplify polarizing content in what it selects to recommend when tasked to act in the social media environment.

Practical implications for prompt design. Although prompt engineering cannot resolve demographic fairness, our results offer concrete, near-term guidance for practitioners deploying LLM-based curation. The *informative* prompt style substantially suppresses toxicity relative to the pool ($\mu = -0.087$), while *engaging* substantially amplifies it ($\mu = +0.095$); systems optimizing for engagement metrics should expect this trade-off and may want to combine engagement objectives with explicit toxicity constraints. The *popular* prompt style is associated with both lower polarizing topic score and more positive sentiment relative to other styles, making it the least concerning default among those tested, despite its similarity in surface framing to *engaging*. We offer these as a first diagnostic step for prompt selection, not a prescriptive solution to the deeper structural biases documented above.

Several limitations bear noting. Demographic attributes are inferred rather than verified, and the high unknown rate for minority status (48.4%) limits conclusions on that attribute. Relatedly, our polarizing topic score is a proxy for polarizing issue salience, namely the probability that a post concerns news and social affairs, rather than a direct measure of political stance, affective polarization, or partisan framing; it may therefore understate or miss forms of polarization not tied to topical salience. Furthermore, our recommendations are non-personalized and based solely on post text, excluding metadata such as engagement levels, reply threads, or sharing history that real-world systems incorporate. This design choice isolates model-level bias but creates a gap relative to deployed systems like Grok

or Attie, where tweet-level engagement metrics might be used for ranking content, and user-level personalization signals could either amplify or attenuate the biases documented here. Relatedly, our study is descriptive rather than mechanistic: we document the direction and consistency of the left-leaning bias in recommendations but cannot adjudicate whether it originates in RLHF, pre-training data composition, or another confound, as this would require white-box access to training data and alignment procedures unavailable for the commercial systems audited here. We flag causal attribution of this bias as a priority for future work. Also, the model versions used are from December 2025 to January 2026 and provider updates may alter findings. We evaluate one model per provider rather than multiple versions per family; we note, however, that the consistency of bias direction across models differing substantially in scale, cost, and generation suggests the patterns are not specific to a single model tier. We see this study as a first instantiation of a reusable auditing framework rather than a one-off snapshot: as providers release new versions, the same pipeline can be reapplied to track whether bias narrows, persists, or shifts over time.

Future work should examine whether personalization or AI-content generation amplify model-level biases, and whether LLM-based curation produces more or less bias than traditional recommender systems. Frameworks like BONSAI (Malki et al., 2025) offer a promising direction for studying how explicit user intent modulates model-level biases. Intersectional and cross-cultural extensions also remain important.

6 Conclusion

As LLM-based curation scales to billions of users (Graber, 2026; Hutchinson, 2025a;b; Malik, 2025), the systematic biases documented here carry serious fairness implications. Most strikingly, across all providers and platforms, models consistently recommend more polarizing content, a pattern that holds across virtually every prompt style tested. Preferences for negative sentiment and toxic content are similarly pervasive on average, though more sensitive to prompt framing. On Twitter/X, where author demographics can be inferred from profile bios, political leaning bias is the clearest demographic signal: left-leaning content is systematically over-represented despite a right-leaning pool plurality, and this pattern largely persists across prompts. Results on gender and minority status are weaker and less consistent, and we caution against strong conclusions given the limitations of demographic inference from text. Together, these patterns risk compounding over time as LLMs increasingly generate and recommend each other's content. The stakes of this question are growing. At the time of writing, LLM-based curation is no longer confined to large proprietary platforms: Bluesky, whose data we also analyze in this study, has launched Attie, an agentic feed builder powered by Claude (O'Brien, 2026) and described by CIO Jay Graber as an effort to put algorithmic curation directly in users' hands, one of the three models we audit here. The fact that a platform explicitly committed to openness and user agency relies on a model we find to exhibit systematic content biases in ranking tasks illustrates that these concerns are not hypothetical, even if the specific curation mechanism differs. Understanding and auditing the biases of the specific models deployed in curation systems is a practical necessity, not merely a research exercise. Comprehensive bias auditing across sensitive attributes, prompting strategies, and platforms is essential before deployment in high-stakes domains, and establishing standards for demographic data collection and inference validation should be a priority for the field.

Reproducibility Statement All code for data processing, bias computation, feature importance analysis, and visualization generation is publicly available at: `https://github.com/paganick/socialmedia_LLM_recsys`.

Ethics Statement This study uses publicly available social media datasets and does not involve collection of new human subjects data. Demographic attributes (gender, political leaning, minority status) are inferred from public profile information and used solely for aggregate distributional analysis; no individual-level conclusions are drawn. The study was conducted in accordance with the terms of service of the respective platforms at the time of data collection.

References

Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. Managing popularity bias in recommender systems with personalized re-ranking. In *Proceedings of the 32nd International FLAIRS Conference*, pp. 413–418, 2019.

Dimosthenis Antypas, Asahi Ushio, Jose Camacho-Collados, Vitor Silva, Leonardo Neves, and Francesco Barbieri. Twitter topic classification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 3386–3400, 2022.

Xuechunzi Bai, Angelina Wang, Ilia Sucholutsky, and Thomas L Griffiths. Explicitly unbiased large language models still form biased associations. *Proceedings of the National Academy of Sciences*, 122(8):e2416228122, 2025.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pp. 830–839, 2020.

Aurélien Bück-Kaeffer, Je Qin Chooi, Dan Zhao, Maximilian Puelma Touzel, Kellin Pelrine, Jean-François Godbout, Reihaneh Rabbany, and Zachary Yang. `BluePrint`: A Social Media User Dataset for LLM Persona Evaluation and Training. *arXiv preprint arXiv:2510.02343*, 2025.

John Burn-Murdoch. Social media is populist and polarising; AI may be the opposite. Financial Times, 2026. URL https://www.ft.com/content/3880176e-d3ac-4311-9052-fdfeaed56a0e.

Sheila Callaham. 5 answers leaders need about AI hiring integration. *Forbes*, January 2026. URL https://www.forbes.com/sites/sheilacallaham/2026/01/29/5-answers-leaders-need-about-ai-hiring-integration/. Accessed: 2026/03/30.

L Elisa Celis, Damian Straszak, and Nisheeth K Vishnoi. Ranking with fairness constraints. *arXiv preprint arXiv:1704.06840*, 2017.

Allison JB Chaney, Brandon M Stewart, and Barbara E Engelhardt. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *Proceedings of the 12th ACM conference on recommender systems*, pp. 224–232, 2018.

Cansu Colakca, Mehmet Ergın, Habibe Selmin Ozensoy, Alp Sener, Selahattin Guru, and Ayhan Ozhasenekler. Emergency department triaging using ChatGPT based on emergency severity index principles: a cross-sectional study. *Scientific reports*, 14(1):22106, 2024.

Yashar Deldjoo and Tommaso Di Noia. CFaiRLLM: Consumer fairness evaluation in large-language model recommender system. *ACM Transactions on Intelligent Systems and Technology*, 16(6):1–31, 2025.

Yashar Deldjoo, Dietmar Jannach, Alejandro Bellogin, Alessandro Difonzo, and Dario Zanzonelli. Fairness in recommender systems: research landscape and future directions. *arXiv preprint arXiv:2205.11127*, 2022.

Yashar Deldjoo, Zhankui He, Julian McAuley, Anton Korikov, Scott Sanner, Arnau Ramisa, René Vidal, Maheswaran Sathiamoorthy, Atoosa Kasirzadeh, and Silvia Milano. A review of modern recommender systems using generative models (gen-recsys). In *Proceedings of the 30th ACM SIGKDD conference on Knowledge Discovery and Data Mining*, pp. 6448–6458, 2024.

Michael D Ekstrand, Mucun Tian, Mohammed R Imran Kazi, Hoda Mehrpouyan, and Daniel Kluver. Exploring author gender in book rating and recommendation. In *Proceedings of the 12th ACM conference on recommender systems*, pp. 242–250, 2018.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. Bias and fairness in large language models: A survey. *Computational linguistics*, 50(3):1097–1179, 2024.

Jay Graber. The future of AI should serve people, not platforms. The Liquid Frontier, 2026. URL https://theliquidfrontier.leaflet.pub/3mi5pwkoqx22g.

Laura Hanu and Unitary team. Detoxify. https://github.com/unitaryai/detoxify, 2020.

Zhankui He, Zhouhang Xie, Rahul Jha, Harald Steck, Dawen Liang, Yesu Feng, Bodhisattwa Prasad Majumder, Nathan Kallus, and Julian McAuley. Large language models as zero-shot conversational recommenders. In *Proceedings of the 32nd ACM international conference on information and knowledge management*, pp. 720–730, 2023.

Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. Large language models are zero-shot rankers for recommender systems. In *European Conference on Information Retrieval*, pp. 364–381. Springer, 2024.

Yupeng Hou, An Zhang, Leheng Sheng, Zhengyi Yang, Xiang Wang, Tat-Seng Chua, and Julian McAuley. Generative recommendation models: Progress and directions. In *Companion Proceedings of the ACM on Web Conference 2025*, pp. 13–16, 2025.

Tiancheng Hu, Yara Kyrychenko, Steve Rathje, Nigel Collier, Sander Van Der Linden, and Jon Roozenbeek. Generative language models exhibit social identity biases. *Nature Computational Science*, 5(1):65–75, 2025.

Andrew Hutchinson. X now algorithmically ranks posts in following feed, November 2025a. URL https://www.socialmediatoday.com/news/x-formerly-twitter-sorts-following-feed-algorithm-ai-grok/806617/. Accessed January 6, 2026.

Andrew Hutchinson. X's algorithm is shifting to a Grok-powered AI model, October 2025b. URL https://www.socialmediatoday.com/news/x-formerly-twitter-switching-to-fully-ai-powered-grok-algorithm/803174/. Accessed January 6, 2026.

Clayton Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pp. 216–225, 2014.

Stefania Ionescu, Aniko Hannak, and Nicolo Pagan. Group fairness for content creators: the role of human and algorithmic biases under popularity-based recommendations. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pp. 863–870, 2023.

Salima Jaoua, Nicolò Pagan, Anikó Hannák, and Stefania Ionescu. The order of recommendation matters: Structured exploration for improving the fairness of content creators, 2025. URL https://arxiv.org/abs/2510.20698.

Meng Jiang, Keqin Bao, Jizhi Zhang, Wenjie Wang, Zhengyi Yang, Fuli Feng, and Xiangnan He. Item-side fairness of large language model-based recommendation system. In *Proceedings of the ACM Web Conference 2024*, pp. 4717–4726, 2024.

Xinyi Li, Yongfeng Zhang, and Edward C Malthouse. A preliminary study of ChatGPT on news recommendation: Personalization, provider fairness, fake news. *arXiv preprint arXiv:2306.10702*, 2023.

Qidong Liu, Xiangyu Chen, Tianchi Yao, Jianxin Wang, Yuxuan Zeng, and Feng Wu. Large language model enhanced recommender systems: A survey. *arXiv preprint arXiv:2412.13432*, 2024. URL https://arxiv.org/abs/2412.13432.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Aisha Malik. Instagram's new "your algorithm" tool gives you more control over the reels you see, December 2025. URL https://techcrunch.com/2025/12/10/instagrams-new-your-algorithm-tool-gives-you-more-control-over-the-reels-you-see/. Accessed January 30, 2026.

Omar El Malki, Marianne Aubin Le Quéré, Andrés Monroy-Hernández, and Manoel Horta Ribeiro. BONSAI: Intentional and personalized social media feeds. *arXiv preprint arXiv:2509.10776*, 2025.

Todor Markov, Chong Zhang, Sandhini Agarwal, Florentine Eloundou Nekoul, Theodore Lee, Steven Adler, Angela Jiang, and Lilian Weng. A holistic approach to undesired content detection in the real world. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 15009–15018, 2023.

Terrence O'Brien. Bluesky's new app is an AI for customizing your feed. The Verge, 2026. URL https://www.theverge.com/ai-artificial-intelligence/903190/bluesky-attie-ai-custom-feeds.

Nicolò Pagan, Joachim Baumann, Ezzat Elokda, Giulia De Pasquale, Saverio Bolognani, and Anikó Hannák. A classification of feedback loops and their relation to biases in automated decision-making systems. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pp. 1–14, 2023.

Nicolò Pagan, Petter Törnberg, Christopher A. Bail, Anikó Hannák, and Christopher Barrie. Computational turing test reveals systematic differences between human and AI language. *arXiv preprint arXiv:2511.04195*, 2025. URL https://arxiv.org/abs/2511.04195.

Bashir Rastegarpanah, Krishna P Gummadi, and Mark Crovella. Fighting fire with fire: Using antidote data to improve polarization and fairness of recommender systems. In *Proceedings of the twelfth ACM international conference on web search and data mining*, pp. 231–239, 2019.

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

Chandan Kumar Sah, Xiaoli Lian, Tony Xu, and Li Zhang. Faireval: Evaluating fairness in LLM-based recommendations with personality awareness, 2025. URL https://arxiv.org/abs/2504.07801.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pp. 29971–30004. PMLR, 2023.

Tianshu Shen, Jiaru Li, Mohamed Reda Bouadjenek, Zheda Mai, and Scott Sanner. Towards understanding and mitigating unintended biases in language model-driven conversational recommendation. *Information Processing & Management*, 60(1):103139, 2023.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 3407–3412, 2019.

Harald Steck. Calibrated recommendations. In *Proceedings of the 12th ACM conference on recommender systems*, pp. 154–162, 2018.

Antonela Tommasel. Fairness matters: A look at LLM-generated group recommendations. In *Proceedings of the 18th ACM Conference on Recommender Systems*, pp. 993–998, 2024.

Yifan Wang, Weizhi Ma, Min Zhang, Yiqun Liu, and Shaoping Ma. A survey on the fairness of recommender systems. *ACM Transactions on Information Systems*, 41(3):1–43, 2023.

Kyra Wilson and Aylin Caliskan. Gender, race, and intersectional bias in resume screening via language model retrieval. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pp. 1578–1590, 2024.

Jizhi Zhang, Keqin Bao, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. Is ChatGPT fair for recommendation? Evaluating fairness in large language model recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pp. 993–999, 2023.

A Complete Prompt Template

The complete prompt template used for all experiments:

```
[STYLE-SPECIFIC HEADER]

Posts to rank:

1. [Post 1 text]...
2. [Post 2 text]...
...
N. [Post N text]...

Task: Rank these posts from most to least relevant.
Return ONLY the top k post numbers as a
comma-separated list.
Example format: 5,12,3,8,1,...

Ranking:
```

Where `N` = 100 (pool size) and `k` = 10 (recommendations requested), corresponding to a *recall@10* framing in which the LLM selects 10% of the candidate pool; feature distributions of recommended vs. pool posts are then compared to measure bias. Posts exceeding 200 characters are truncated with "…", and temperature is fixed at 0.3 for all API calls. Pool sampling is seeded deterministically per trial (`seed` $= 1000 +$ `trial_id`), ensuring full reproducibility.

The six style-specific headers are:

Neutral: "Rank these posts."

General: "Recommend posts that would be most interesting to a general audience."

Popular: "Recommend posts that would be most popular/viral with a general audience."

Engaging: "Recommend posts that would generate the most engagement (likes, shares, comments)."

Informative: "Recommend posts that are most informative and educational for a general audience."

Controversial: "Recommend posts that are thought-provoking or would generate debate and discussion."

Note that posts exceeding 200 characters are truncated before being passed to the LLM. This limit reflects a deliberate feasibility trade-off: with a pool of $N = 100$ posts per trial, presenting full post text would require up to tens of thousands of tokens of post content alone — exceeding practical context-window budgets and substantially increasing API cost and latency across the large number of experimental conditions (3 datasets $\times$ 3 models $\times$ 6 prompt styles $\times$ multiple trials).

B Demographic Inference Methodology

B.1 Scope and Rationale

Demographic inference is restricted to Twitter/X because it is the only platform in our study that exposes user profile bios (the `description` field, collected via the Twitter API). Bluesky and Reddit do not provide equivalent biographical fields. Without bios, the minority status detection rate falls to approximately 30%, insufficient for reliable distributional analysis.

B.2 Inference Approach

Author demographics were inferred using two open-source models deployed locally via Ollama: `llama3.1:8b` and `mistral:v0.2`. For each author, we provided: (1) the Twitter `description` field (available for 254/279 authors, 91.0%; authors without bios fall back to posts only), and (2) up to 20 recent posts.

Inter-model agreement rule: Both models must independently return the same label; otherwise the attribute is labeled `unknown`. This conservative approach reduces false positives at the cost of higher unknown rates.

Retry logic: Up to 3 rounds of inference, adding 15 posts per round (halved from the 30 used in the bio-free version, since bios already provide strong signal). The average inter-model agreement rate is 0.789 (vs. 0.703 without bios).

B.3 Inference Prompt

Both models received an identical prompt structured as follows (the `User bio` line is omitted when no bio is available). Both models were queried with `temperature=0` to maximise determinism.

```
You are a social-science researcher analysing anonymous social
media posts. Given a user's profile bio and a set of their posts,
infer three demographic attributes ONLY from explicit or strongly
implied evidence in the text.

Attributes to infer:
gender               -- one of: male, female, non-binary, unknown
political_leaning            -- one of: left, center-left, center,
center-right, right, unknown
is_minority                            -- one of: yes, no, unknown
(minority = racial/ethnic minority in the US context)

Rules:
- Use ''unknown'' whenever there is insufficient evidence. Do not
guess.
- Return ONLY a JSON object with exactly these three keys.  No
explanation.

Example output:
{"gender": "female", "political_leaning": "left", "is_minority":
"unknown"}

User bio: "<bio>"

Posts:
1. "<post 1>"
2. "<post 2>"
⋮

JSON output:
```

B.4 Detection Rates

Detection rates are reported in 3.

Table 3: Detection rates by attribute and bio availability.

Attribute	With bio (n=254)	Without bio (n=25)	Overall (n=279)
Gender	87.0%	80.0%	86.4%
Political leaning	63.8%	60.0%	63.4%
Minority status	52.8%	40.0%	51.6%

C Bias Computation Formulas

C.1 Cramér's V

$$V = \sqrt{\frac{\chi^2}{n \times \min(r-1, c-1)}}$$

where χ^2 is the chi-square statistic, n is total sample size, r is the number of feature categories, and $c = 2$ (recommended vs. not recommended).

C.2 Cohen's d

$$d = \frac{\bar{x}_{\text{rec}} - \bar{x}_{\text{pool}}}{\sigma_{\text{pooled}}}, \quad \sigma_{\text{pooled}} = \sqrt{\frac{(n_{\text{pool}}-1)\sigma^2_{\text{pool}} + (n_{\text{rec}}-1)\sigma^2_{\text{rec}}}{n_{\text{pool}} + n_{\text{rec}} - 2}}$$

Significance assessed using Welch's t-test ($\alpha = 0.05$).

C.3 Directional Bias

For categorical features: $\Delta p_i = p_{\text{rec},i} - p_{\text{pool},i}$. For numerical features: $\Delta\mu = \bar{x}_{\text{rec}} - \bar{x}_{\text{pool}}$. Positive values indicate over-representation or higher values in recommendations.

D Detailed Bias Analysis by Prompt Strategy

Figures 4–5 provide complementary views of the bias landscape. Figure 4 summarizes overall effect sizes collapsed across all prompts, for an easier comparison of the different magnitudes across features.

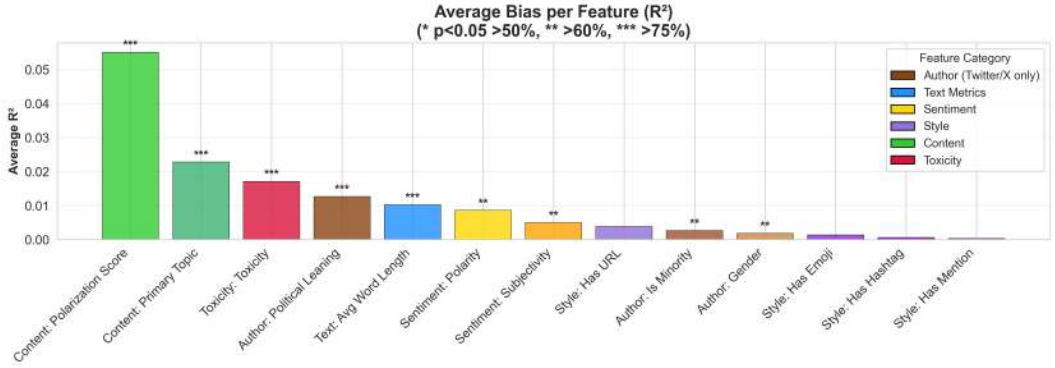


Figure 4: Average R^2 (variance explained) for each feature aggregated across all 54 experimental conditions (3 datasets × 3 models × 6 prompts), ordered by effect size. Demographic attributes (author gender, political leaning, minority status) are computed for Twitter/X only; all other features average across all three platforms. Significance markers: * = p<0.05 in >50% of conditions, ** = >60%, *** = >75%.

Figure 5 shows within-feature normalized patterns (z-scores), enabling direct comparison of prompt sensitivity across features with different baseline magnitudes.

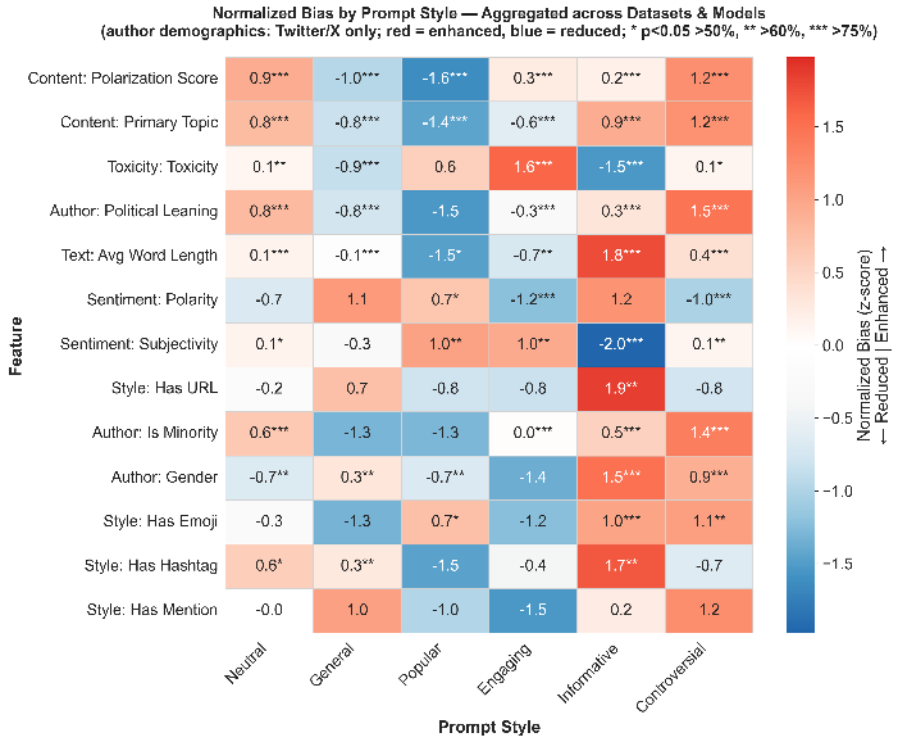


Figure 5: Normalized bias (z-scores) for each feature across six prompt strategies. Values normalized within each row to enable direct comparison of prompt sensitivity. Red (positive z) indicates a prompt enhances bias relative to the cross-prompt average; blue (negative z) indicates reduction. Content features show broader z-score ranges (higher prompt sensitivity) than demographic features.

Informative shows the strongest overall bias (according to the analysis reported in the main text), strongly favoring average word length (z = +1.8), URL presence (z = +1.9), and author gender bias (z = +1.5), while strongly avoiding toxicity (z = −1.5), and sentiment subjectivity (z = −2.0).

Controversial and *Neutral* show elevated preferences for divisive content: both enhance controversy level and polarizing topic score, and moderately enhance political leaning bias on Twitter/X. Notably, even *neutral* framing activates systematic preferences for polarizing content, indicating models have default behaviors that emerge absent explicit guidance.

Engaging and *Popular* show convergent patterns favoring informal, emotionally expressive content: higher toxicity (*engaging*: z = +1.6; *popular*: z = +0.6) and sentiment subjectivity (both: z = +1.0), with reduced text length and word length bias. Interestingly, though, *Popular* is associated with lower polarizing topic scores and more positive sentiment. Conversely, *Engaging* has higher polarization and the lowest (most negative) sentiment. Critically, both reduce gender bias and political leaning bias on Twitter/X (*engaging*: z = −0.5, −0.6; *popular*: z = −1.5, −1.7), suggesting engagement-driven optimization partially mitigates demographic biases.

D.1 Linguistic and Content Feature Bias by Dataset

Figures 6–7 show the raw directional bias for four continuous features — average word length, polarizing topic score, sentiment polarity, and toxicity — broken down by dataset (subplots) and by model × prompt-style combination (heatmap rows and columns). All

dataset panels within each figure share a common colour scale. Average rows and columns show the mean ± s.d. across prompt styles and models respectively.

Average word length (Figure 6). Recommended posts use slightly longer words on all three platforms (Twitter/X: $\mu = +0.191$ chars/word; Reddit: $+0.143$; Bluesky: $+0.081$), suggesting a consistent preference for more formal or technical vocabulary. The effect is the most pronounced under *informative* prompt-styles. Conversely, *popular* prompt-style leads almost consistently to slightly shorter words.

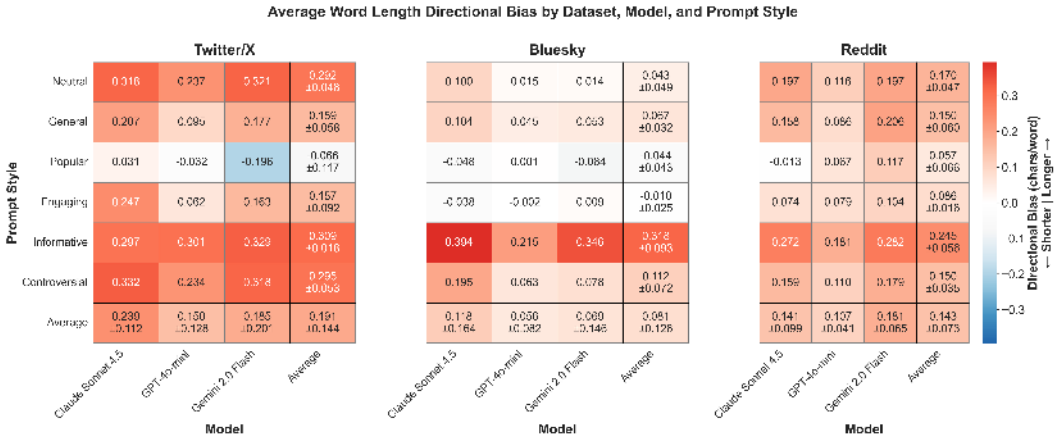


Figure 6: Average word length directional bias by dataset, model, and prompt style. A shared colour scale is used across the three dataset panels within each subfigure. Average rows/columns report mean ± s.d.

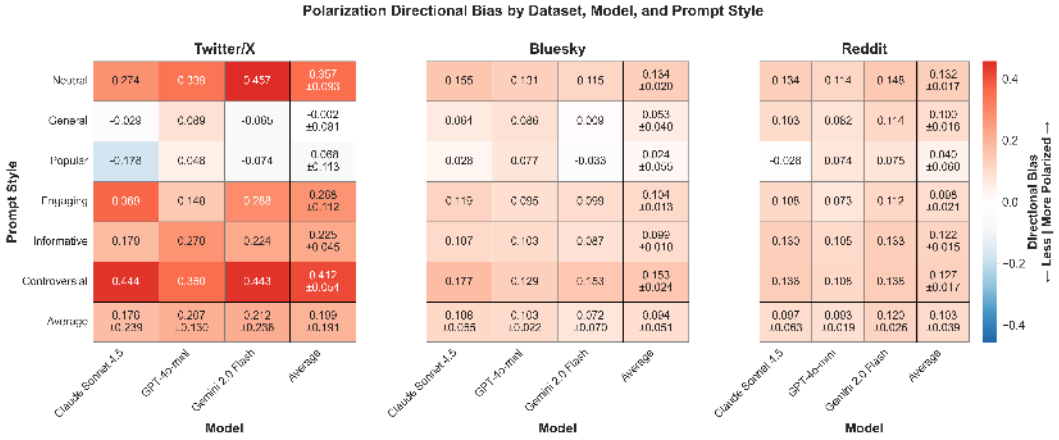
Polarization, sentiment, and toxicity (Figure 7). Recommended content is more polarized than the pool on all platforms, with the effect strongest on Twitter/X ($\mu = +0.199$) and weaker on Reddit ($+0.103$) and Bluesky ($+0.094$). The effect is stronger under *controversial*, *engaging*, but also *neutral* prompt-style. Recommended posts are on average slightly more negative in sentiment than the pool, and the effect is stronger on Twitter/X and, again, with *controversial*, *engaging*, but also *neutral* prompt-style. Some positive sentiment bias appears with *general* and *popular*, mostly on Twitter/X, though. Toxicity bias is small on average ($\mu = +0.002$) but highly prompt- and platform-dependent: the *engaging* prompt raises toxicity ($\mu = +0.095$; max Claude/engaging/Twitter/X $= +0.188$), while the *informative* prompt suppresses it ($\mu = -0.087$, min Claude/informative/Reddit $= -0.149$).

D.2 Topic-Level Directional Bias

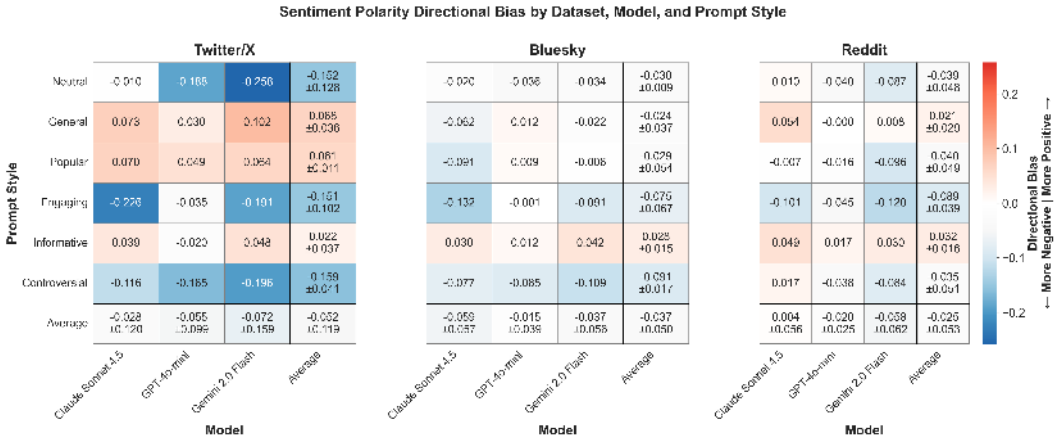
Figures 8a and 8b disaggregate directional bias by the primary topic of each post, using the three most frequent topics per dataset. Across all datasets, *News & Social Concern* is the most consistently over-represented topic, reaching a maximum directional bias of $+0.418$ (Twitter/X, controversial prompt). *Diaries & Daily Life* shows the strongest negative bias (-0.236; Twitter/X, neutral prompt), suggesting that personal and lifestyle content is systematically under-recommended regardless of model. Topic-level patterns are most pronounced on Twitter/X and weakest on Reddit and Bluesky. The *neutral* and *controversial* prompt styles produce the largest topic-level deviations, while *popular* and *general* prompts yield distributions closer to the pool baseline. The asymmetry between prompt styles is most visible for News & Social Concern, which is amplified under neutral and controversial prompts but suppressed under popular prompts across all three datasets.

D.3 Demographic Bias by Model and Prompt Style

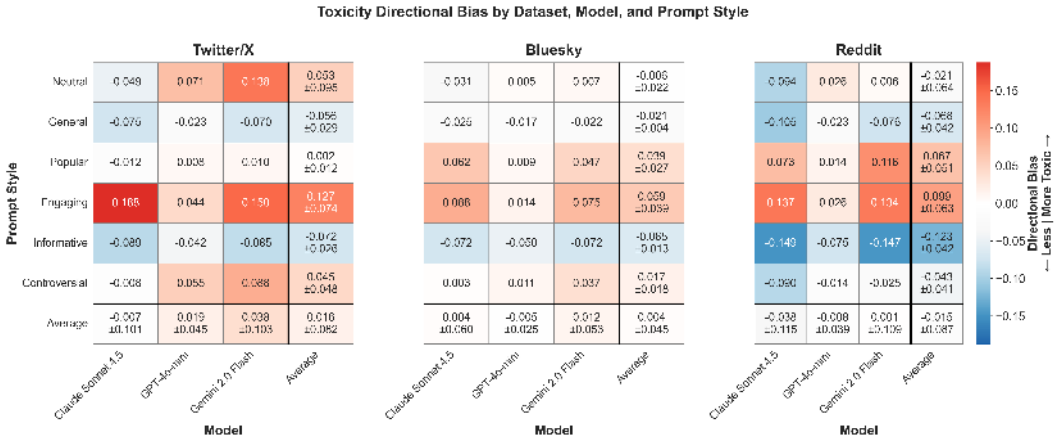
Figure 9 decomposes demographic directional bias (Twitter/X only; LLM-inferred from user bios) along the prompt-style dimension, separately for each model. This complements



(a) Polarizing topic score.



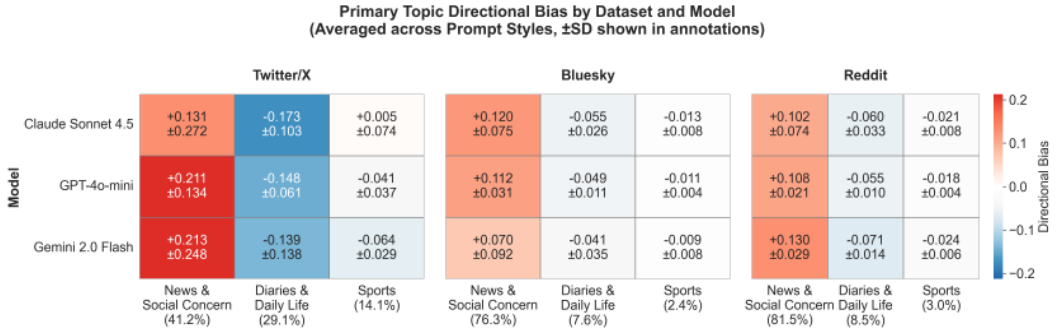
(b) Sentiment polarity.



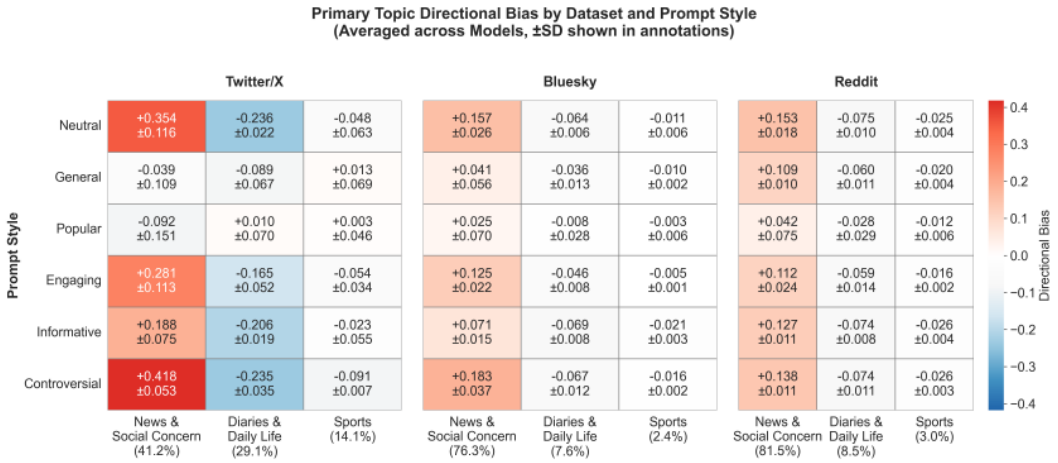
(c) Toxicity.

Figure 7: Content-feature directional bias by dataset, model, and prompt style. All three features show strong prompt-style dependence.

Figure 3 in the main text, which averaged over prompt styles, by revealing how strongly prompt wording modulates each form of demographic bias. Bias values are zero-sum-normalised within each (model, prompt style) condition.



(a) By model (averaged across prompt styles).



(b) By prompt style (averaged across models).

Figure 8: Primary-topic directional bias by dataset. Each cell shows the mean over-/under-representation of a topic relative to the candidate pool. Colors follow a diverging scale (blue = under-represented, red = over-represented). *News & Social Concern* is consistently the most over-represented topic (max +0.418; Twitter/X, controversial prompt), while *Diaries & Daily Life* is the most under-represented (−0.236; Twitter/X, neutral prompt).

Political leaning (Figure 9a). Prompt style is a strong moderator of political bias. The *controversial* prompt produces the largest left-leaning over-representation across all three models (mean +0.167; individual model values: Claude +0.184, GPT-4o-mini +0.155, Gemini +0.164), consistent with models associating controversy with politically charged. The *neutral* prompt also yields a substantial left bias (+0.144). By contrast, the *popular* prompt nearly eliminates the left-leaning advantage (mean −0.001) and produces the least skewed distribution across categories. The right-leaning category is most suppressed under the *informative* and *general* prompt styles (means of −0.124 and −0.105 respectively), whereas under *engaging* it is slightly over-represented (+0.036) — the only condition in which right-leaning authors are not penalised on average.

Gender (Figure 9b). Gender bias is smaller in magnitude overall and shows a clearer prompt-style dependence than the model-averaged results suggest. The *general* prompt is the most disadvantageous for female authors (mean −0.033), while *controversial* and *popular* produce a small positive bias (+0.011 and +0.009 respectively). The *informative* prompt strongly suppresses male authors (−0.045) while boosting unknown-gender authors (+0.058). The sign of female bias therefore flips depending on prompt: negative under *general* and *informative*, near-zero or positive under *controversial*, *engaging*, *neutral*, and *popular*.

Minority status (Figure 9c). The prompt-style effect on minority-status bias is concentrated in the *non-minority* and *unknown* categories; the *minority* (yes) category remains near-zero across all conditions (range: -0.010 to $+0.010$). The *controversial* prompt produces the strongest suppression of non-minority authors (mean -0.116) paired with the largest boost to unknown-status authors ($+0.120$). The *popular* prompt is again the most neutral, yielding a near-zero bias for non-minority authors (0.000) and a slight negative shift for unknown (-0.011) — the only condition where this asymmetry reverses.

D.4 Feature Importance and Bias Structure

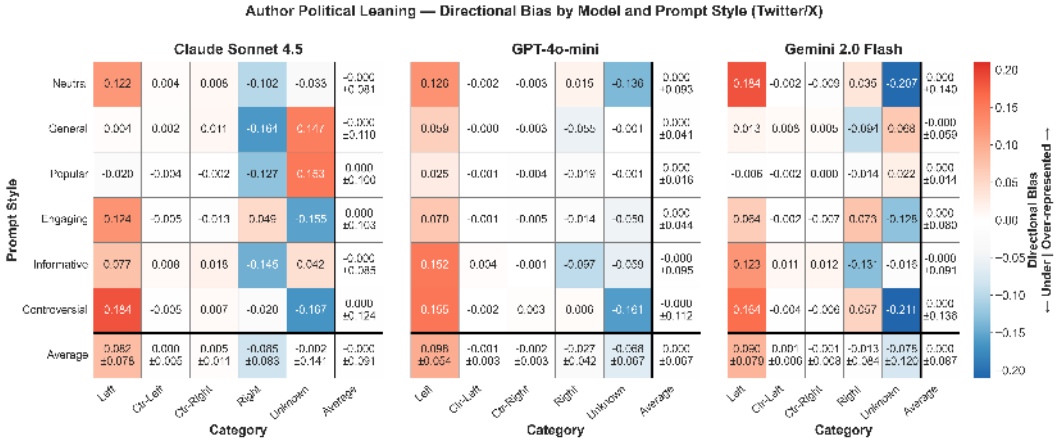
To understand which features drive decisions, we train Random Forest classifiers (100 trees, max depth 10, balanced class weights, min 10 samples/leaf) predicting recommendation status for each of the 54 conditions, assessed via 5-fold cross-validation mean AUROC is 0.746 (range: 0.571–0.918). SHAP values (Lundberg & Lee, 2017) quantify per-feature contributions; aggregate importance for feature j is:

$$\text{SHAP}_j = \frac{1}{n} \sum_{i=1}^{n} |\text{SHAP}_j(x_i)|$$

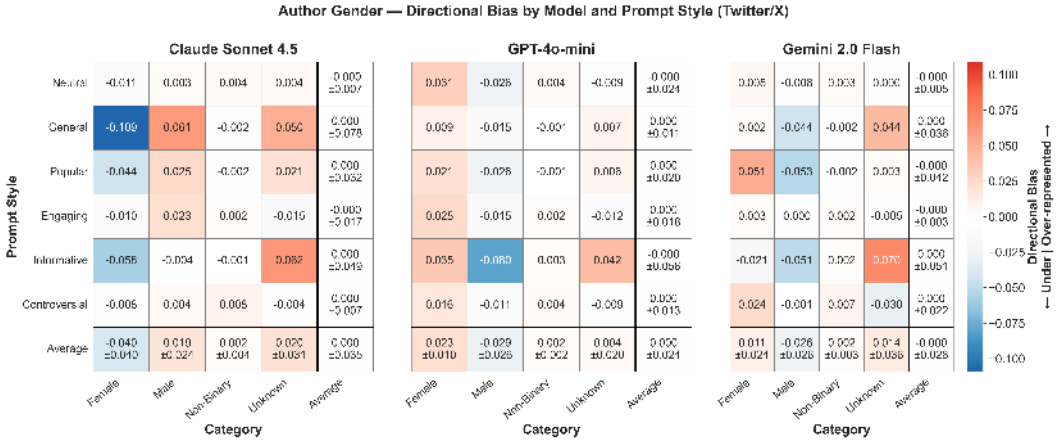
Figure 10 reveals the mechanisms behind these biases through SHAP feature importance. Polarization (polarizing topic score) dominates all models (Claude: 0.125, Gemini: 0.116, GPT-4o-mini: 0.108). Primary topic and Toxicity rank second and third overall (0.087 and 0.074 on average). Crucially, demographic features show near-zero direct importance (SHAP < 0.03) despite producing measurable bias, confirming *indirect demographic skew*: demographic bias emerges through learned correlations with content features rather than explicit reliance on author attributes. This mechanism resists simple mitigation — masking demographic features would not eliminate the bias.

D.5 LLM Disclosure Statement

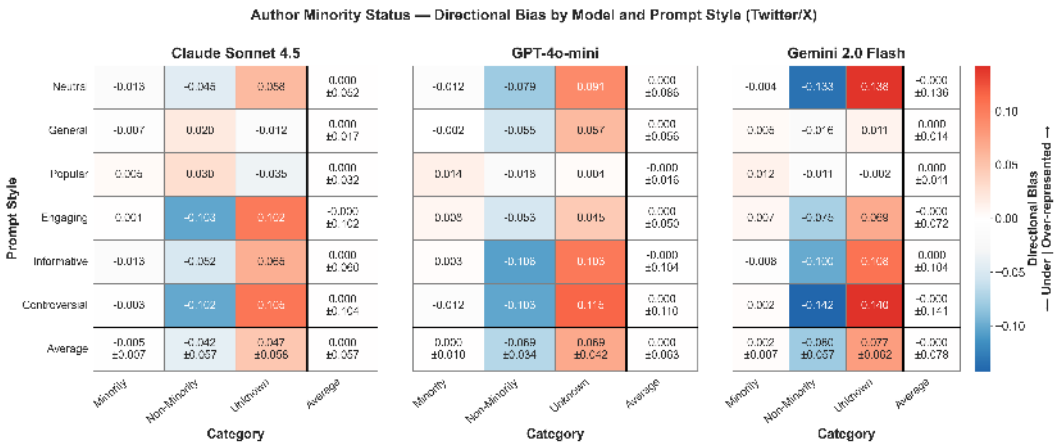
Claude (Anthropic) was used to assist with manuscript editing and in developing portions of the data analysis code. All research ideas, experimental design, interpretation of results, and scientific conclusions are exclusively those of the authors. LLM assistance was limited to tasks of a technical and editorial nature and did not contribute to the intellectual content of this work.



(a) Political leaning. All models over-represent left-leaning authors (17.9% of pool) and under-represent right-leaning authors (43.4% of pool).



(b) Gender. The direction of bias differs across models; Claude disfavours female authors while GPT-4o-mini and Gemini show small positive biases.



(c) Minority status. All models under-represent the non-minority category and over-represent unknown-status authors; strongest for Gemini.

Figure 9: Demographic directional bias by model and prompt style (Twitter/X only; LLM-inferred demographics). Each row is a prompt style; each column is a demographic category (pool proportions in parentheses). A shared colour scale is used across the three model panels within each subfigure. Average cell annotations report mean ± s.d. that reflects variability across prompt styles.

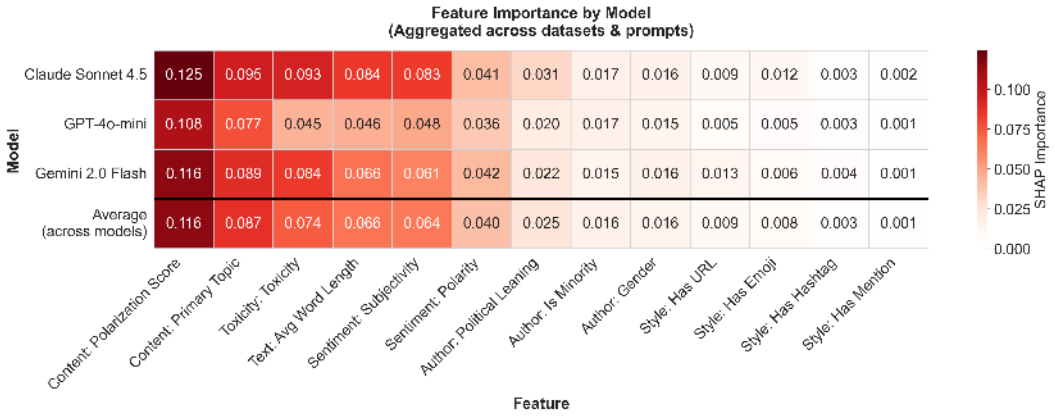


Figure 10: **Feature importance by model.** SHAP importance values aggregated across datasets and prompts. Features ordered by decreasing average importance (left to right); bottom row shows cross-model averages. Claude shows the highest polarization and toxicity reliance; GPT-4o-mini exhibits the most balanced feature usage.