Evaluating LLM-Based Bias Detection on Wikipedia: A Prompt Strategy Analysis Using NPOV Guidelines

Ashik Ahamed
ahameda@clarkson.edu
Clarkson University
Potsdam, New York, USA

Max Wang
mwang@clarkson.edu
Clarkson University
Potsdam, New York, USA

Jeanna Matthews
jnm@clarkson.edu
Clarkson University
Potsdam, New York, USA

Abstract

Wikipedia's Neutral Point of View (NPOV) policy is central to maintaining the integrity of one of the world's most widely accessed information sources. As AI-powered moderation tools become increasingly prevalent in adaptive web environments, understanding their reliability and limitations is critical. This paper investigates the effectiveness of Large Language Models (LLMs) in detecting neutrality violations in Wikipedia text using the Wikimedia Neutrality Corpus (WNC). We evaluate three Claude models—`claude-3-haiku`, `claude-haiku-4-5`, and `claude-sonnet-4-5`—across four prompt strategies ranging from direct instruction to policy-guided few-shot prompting. Performance is further analyzed across topical domains, macro-topic categories, bias types (framing, demographic, and epistemological), and discourse communities. Our results show that `claude-sonnet-4-5` with NPOV guidelines and detailed explanations achieves the highest accuracy of 69.0%, yet performance varies substantially across domains—medical discourse showing the highest disagreement rate at 64.3%. These findings reveal important risks and limitations of deploying LLMs for automated content moderation on the adaptive web, and offer practical guidance for prompt design in policy-grounded classification tasks.

CCS Concepts

• **Information systems** → **Sentiment analysis**; • **Computing methodologies** → *Natural language processing*; • **Social and professional topics** → User characteristics.

Keywords

bias detection, Wikipedia, NPOV, large language models, prompt engineering, natural language processing, AI-fueled personalization, content moderation, adaptive web

ACM Reference Format:
Ashik Ahamed, Max Wang, and Jeanna Matthews. 2026. Evaluating LLM-Based Bias Detection on Wikipedia: A Prompt Strategy Analysis Using NPOV Guidelines. In *18th ACM Web Science Conference Companion (WebSci Companion '26), May 26–29, 2026, Braunschweig, Germany.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3795513.3810456

1 Introduction

Wikipedia is among the most visited websites in the world and serves as a primary source of information for hundreds of millions of people daily. Its Neutral Point of View (NPOV) policy requires that articles represent all significant viewpoints fairly and without advocacy, making neutrality a cornerstone of the platform's credibility and usefulness. Maintaining NPOV compliance at the scale of over 60 million articles across hundreds of language editions is, however, a formidable challenge. Wikipedia depends on a large and distributed community of volunteer editors to identify and revise biased content—a process that is labor-intensive, inconsistent across contributors, and difficult to scale [3]. The complexity of this ecosystem is further reflected in the nuanced spectrum between human edits and automated bot contributions, which introduces additional variability in content quality and tagging practices [1].

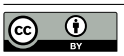
The rise of Large Language Models (LLMs) has created new opportunities for automating content quality tasks at scale. These models can process natural language with remarkable fluency and, through appropriate prompting, can be directed to apply complex editorial policies such as NPOV. Yet the deployment of such systems in live platforms raises important questions about reliability, fairness, and unintended societal effects. In the context of the adaptive web—where recommender systems, personalization engines, and AI-driven moderation tools increasingly shape the information users encounter—the stakes are especially high. An automated bias detection system that consistently mislabels neutral expert content as biased, or fails to detect genuinely biased statements in sensitive domains, could distort knowledge access for large populations.

This work is framed within the ABIS 2026 theme of *AI-fueled personalization and its societal effects.* As LLMs are increasingly integrated into the adaptive web as intelligent content filters and moderation agents, it is critical to evaluate both their capabilities and their failure modes systematically. Our study contributes to this evaluation by examining how three generations of Claude models perform on the NPOV bias detection task under four distinct prompt strategies, and by analyzing where and why these models fail across topical domains and discourse communities.

We use the Wikimedia Neutrality Corpus (WNC) [5] as our evaluation benchmark. The WNC provides sentence-level pairs of biased Wikipedia text and human-revised neutral versions, enabling supervised classification experiments grounded in real-world editorial practice. We evaluate models across 400 balanced sentences, stratified by topic and discourse community, to produce a nuanced picture of model performance.

Our main contributions are:

- A comparative evaluation of three Claude model generations on Wikipedia NPOV bias detection using zero-shot and few-shot prompting.
- A systematic analysis of four prompt strategies—from direct instruction to policy-guided few-shot prompting—showing



WebSci Companion '26, Braunschweig, Germany

ACM ISBN 979-8-4007-2492-3/2026/05
https://doi.org/10.1145/3795513.3810456

how prompt design dramatically affects classification performance.
- A topic-aware analysis across micro and macro topic categories, identifying domains where model performance is strongest and weakest.
- A bias-type distribution analysis (framing, epistemological, demographic) conducted as a secondary experiment on the 201 human-labeled biased sentences, using four dedicated prompts of increasing richness across all three Claude models, showing how prompt context shapes detection of surface-visible versus subtle bias types.
- A discourse-community analysis demonstrating that professional language style is a key driver of model error, with medical discourse presenting the highest challenge at 64.3% disagreement.
- Practical recommendations for deploying LLMs in adaptive web content moderation pipelines.

2 Related Work

2.1 Bias Detection in Wikipedia

Research on Wikipedia bias spans multiple dimensions including ideological bias, gender bias, and systemic coverage gaps [3]. Early work relied on lexicon-based approaches, identifying bias through the presence of subjective or loaded language. More recent work applied supervised machine learning to classify biased sentences using the WNC dataset [5], which provides human-annotated revision pairs and has become the standard benchmark for Wikipedia NPOV research. These approaches achieve moderate accuracy but require large amounts of labeled training data and struggle to generalize across topic domains.

2.2 LLMs for Content Moderation

LLMs have demonstrated strong performance on text classification tasks including toxicity detection, hate speech identification, and misinformation classification [7]. Brown et al. [2] showed that large models can perform complex classification in zero-shot and few-shot settings by providing appropriate instructions in the prompt—without task-specific fine-tuning. More recent work has applied instruction-tuned LLMs to content policy enforcement, finding that alignment with specific policy guidelines improves classification consistency and reduces error rates. However, relatively little work has systematically compared prompt strategies for editorial neutrality tasks specifically, where the target policy is highly nuanced and domain-sensitive.

2.3 Prompt Engineering

Prompt engineering has emerged as a foundational skill for deploying LLMs effectively [6]. Chain-of-thought prompting encourages step-by-step reasoning before producing an answer, improving performance on complex tasks. Policy-enriched prompts have been found to improve model calibration and reduce hallucination. Our four-prompt strategy builds on these findings, progressively enriching the prompt from direct instruction through policy definition, detailed explanation, and illustrated examples—providing a controlled comparison of how each layer of context affects performance.

2.4 Adaptive Web and Societal Effects

The deployment of AI systems on the adaptive web—including recommender systems, search engines, and content moderation tools—has been linked to filter bubbles, information asymmetry, and algorithmic amplification of existing biases [4]. Automated moderation systems can encode and propagate biases present in their training data or decision logic. Our work contributes empirical evidence that LLM-based moderation performance is domain-sensitive and that errors are systematically concentrated in specialized professional discourse communities, which has important implications for equitable information access.

3 Dataset and Methodology

3.1 Wikimedia Neutrality Corpus

The Wikimedia Neutrality Corpus (WNC) [5] is a large-scale dataset of Wikipedia revision pairs in which volunteer editors revised biased sentences to comply with the NPOV policy. Each record includes the original biased sentence (`src_raw`), the human-revised neutral sentence (`tgt_raw`), and metadata including a revision ID and topic probability scores. The corpus covers a wide range of topics and writing styles, reflecting the diversity of Wikipedia content across subject areas. Unlike supervised approaches that require task-specific fine-tuning on labeled WNC training data, this study evaluates zero-shot and few-shot prompting, addressing the practically relevant question of whether LLMs can enforce NPOV policy without any labeled training data.

3.2 Data Processing and Balancing

We filtered the `biased.full` split to retain only instances where `src_raw` differs from `tgt_raw`, ensuring only genuine content modifications are included. All such instances were labeled *biased*. The neutral dataset was retained in full, with all sentences labeled *neutral*. Both datasets were standardized to include revision ID, sentence text, and human label, then merged into a unified evaluation set.

To prevent class imbalance from distorting evaluation metrics, we constructed a balanced dataset by randomly sampling an equal number of biased and neutral sentences based on the size of the smaller class. The combined dataset was shuffled using a fixed random seed of 42 for full reproducibility. A total of **400 sentences** (201 biased, 199 neutral) were used across all experiments. Topic assignment was performed by selecting each sentence's dominant topic based on the highest topic probability score in the corpus metadata, enabling topic-stratified analysis.

3.3 Models Evaluated

We evaluated three Claude models accessed via the Anthropic API:

- `claude-3-haiku-20240307`: A lightweight model from the Claude 3 generation.
- `claude-haiku-4-5-20251001`: An updated Haiku model from the Claude 4.5 generation with improved instruction following.
- `claude-sonnet-4-5-20250929`: A mid-tier model from the Claude 4.5 generation, representing the strongest model tested.

For each sentence, the model was prompted to return a single-word classification (*biased* or *neutral*). Retry logic was implemented to handle transient API failures. All experiments were conducted at default temperature settings to ensure consistency.

3.4 Prompt Strategies

Four prompt strategies of increasing richness were evaluated:

Prompt 1 — Direct Instruction (P1): Only a direct classification instruction is provided, with no policy context or persona framing. The omission of the "Act as an expert Wikipedia editor" instruction is intentional; P1 is designed as a clean zero-shot baseline that avoids priming the model with any Wikipedia-specific context, allowing the contribution of persona and policy framing in subsequent prompts to be isolated. This serves as the zero-shot baseline.

Prompt 2 — NPOV Guidelines (P2): The model receives a concise statement of Wikipedia's NPOV policy before the classification instruction. The condensed bullet format in P2, without the four itemized principles present in P3 and P4, is intentional; this progressive elaboration of policy structure across prompts allows the effect of increasing policy detail to be measured incrementally. This tests whether high-level policy awareness improves performance.

Prompt 3 — NPOV Guidelines + Explanation (P3): The model receives the NPOV policy accompanied by detailed explanations of each principle, including avoiding loaded language, representing minority views fairly, and not stating opinions as facts. This is the richest zero-shot prompt.

Prompt 4 — NPOV Guidelines + Explanation + Examples (P4): Building on P3, the model also receives illustrative examples of biased and neutral sentences demonstrating each NPOV principle. This constitutes a few-shot prompting strategy.

3.5 Evaluation

For each model and prompt, we compute accuracy, precision, recall, and F1-score using human annotations as ground truth. We analyze confusion matrices and false positive/negative rates per configuration.

3.6 Bias Type Classification

Bias type classification was conducted as a secondary experiment entirely independent of the main P1–P4 neutrality detection task. The 201 human-labeled biased sentences were passed through four dedicated bias-type detection prompts of increasing richness: direct label instruction (BT-Direct), definition only (BT-Def), definition with explanation (BT-Def+Expl), and definition with explanation and examples (BT-Def+Expl+Ex). All three Claude models were evaluated under each bias-type prompt condition. Each sentence was classified into one of four categories: framing, epistemological, demographic, or none. Results are reported descriptively to examine how prompt richness shapes the model's ability to distinguish between surface-visible and subtle bias types.

4 Results

4.1 Overall Accuracy Across Models and Prompts

Table 1 presents accuracy across all model and prompt combinations tested.

Table 1: Accuracy (%) by Model and Prompt Strategy (n=400)

Model	P1	P2	P3	P4
claude-3-haiku	58.00	56.00	56.50	60.00
claude-haiku-4-5	63.00	65.00	65.50	63.50
claude-sonnet-4-5	51.75	67.75	**69.00**	67.00

The best-performing configuration is `claude-sonnet-4-5` with Prompt 3 (69.0%). A striking finding is the near-random performance of `claude-sonnet-4-5` on P1 (51.75%): without policy context, this model predicted *neutral* for 378 of 400 sentences, effectively failing to detect bias. Adding NPOV guidelines and explanations raises accuracy sharply to 69.0%, an improvement of over 17 percentage points. Notably, `claude-3-haiku` shows an apparent inverse pattern where P1 (58.00%) outperforms both P2 (56.00%) and P3 (56.50%); however, McNemar's test confirms this difference is not statistically significant ($p = 0.792$), indicating the pattern reflects noise rather than a genuine sensitivity to prompt enrichment. `claude-haiku-4-5` shows the most stable behavior across all prompts (63–65.5%). To assess statistical significance across all key comparisons, we applied McNemar's test to sentence-level predictions. For `claude-sonnet-4-5`, the P1-to-P3 improvement is strongly significant (statistic = 22.55, $p < 0.001$), confirming the 17-percentage-point gain is not due to chance. The P3-to-P4 drop for `claude-sonnet-4-5` is not statistically significant (statistic = 1.94, $p = 0.164$), suggesting that while examples do not improve performance, they do not reliably hurt it either — making P3 the recommended configuration. For `claude-3-haiku`, the P2-to-P4 improvement is significant (statistic = 5.83, $p = 0.016$), suggesting few-shot examples benefit smaller models. In contrast, all `claude-haiku-4-5` comparisons remain non-significant ($p > 0.05$), indicating this model is least sensitive to prompt design changes.

4.2 Full Classification Metrics

Table 2 presents precision, recall, and F1-score for all 12 configurations.

4.3 Confusion Matrix: Best Configuration

Table 3 shows the confusion matrix for `claude-sonnet-4-5`, P3.

The model missed bias in 64 cases (false negatives, 31.8%) and over-flagged neutral content in 60 cases (false positives, 30.2%). This roughly symmetric error distribution is a positive property for a moderation tool, as it indicates the model is not systematically skewed toward either label.

Table 2: Full Classification Metrics by Model and Prompt Strategy

Model	Prompt	Accuracy	Precision	Recall	F1
claude-3-haiku	P1 Direct	58.00%	57.82%	60.70%	59.22%
claude-3-haiku	P2 NPOV	56.00%	53.54%	94.03%	68.23%
claude-3-haiku	P3 NPOV+Expl	56.50%	54.10%	88.56%	67.17%
claude-3-haiku	P4 NPOV+Ex+Expl	60.00%	57.62%	77.11%	65.96%
claude-haiku-4-5	P1 Direct	63.00%	71.90%	43.28%	54.04%
claude-haiku-4-5	P2 NPOV	65.00%	62.66%	75.12%	68.33%
claude-haiku-4-5	P3 NPOV+Expl	65.50%	64.79%	68.66%	66.67%
claude-haiku-4-5	P4 NPOV+Ex+Expl	63.50%	61.70%	72.14%	66.51%
claude-sonnet-4-5	P1 Direct	51.75%	68.18%	7.46%	13.45%
claude-sonnet-4-5	P2 NPOV	67.75%	71.69%	59.20%	64.85%
claude-sonnet-4-5	P3 NPOV+Expl	**69.00%**	**69.54%**	**68.16%**	**68.84%**
claude-sonnet-4-5	P4 NPOV+Ex+Expl	67.00%	68.85%	62.69%	65.62%

Table 3: Confusion Matrix: claude-sonnet-4-5, Prompt 3

	Pred. Biased	Pred. Neutral
Human Biased (201)	137	64
Human Neutral (199)	60	139

5 Topic and Bias Type Analysis

5.1 Macro-Topic Performance

Table 4 reports classification metrics across four macro-topic categories with sufficient representation in the 400-sentence evaluation set (both biased and neutral sentences), evaluated using the best-performing configuration from Section 4 — `claude-sonnet-4-5` with P3. Topic assignment was performed automatically using the dominant topic probability scores provided in the WNC corpus metadata. Of the 400 sentences, human-labeled biased sentences were distributed as follows: Culture (34 of 60), Geography (15 of 30), History & Society (13 of 28), and STEM (5 of 8); the model predicted bias in Culture (26 of 60), Geography (9 of 30), History & Society (12 of 28), and STEM (3 of 8), consistently under-predicting bias relative to human labels across all topics. The remaining sentences belonged to smaller topic categories or lacked reliable topic assignments and are not shown individually.

Table 4: Macro-Topic Performance: claude-sonnet-4-5, P3

Topic	N	Acc.	Prec.	Recall
Culture	60	70.0%	80.8%	61.8%
Geography	30	66.7%	77.8%	46.7%
History & Society	28	67.9%	66.7%	61.5%
STEM	8	75.0%	100.0%	60.0%

Culture achieves the highest accuracy (70.0%), likely because cultural bias often manifests as explicit evaluative language. Geography shows the lowest recall (46.7%), suggesting geographic framing bias—e.g., subtle characterizations of places, regions, or peoples—is harder to detect. STEM topics yield perfect precision (100%) but only moderate recall (60%), indicating the model is appropriately conservative in labeling scientific content as biased, but misses genuine violations.

5.2 Bias Type Distribution by Prompt

Table 5: Bias Type Distribution by Bias-Type Detection Prompt (n=201). Note: these prompts are distinct from the P1–P4 neutrality classification prompts reported in Tables 1 and 2

Bias Type	BT-Direct	BT-Def	BT-Def+Expl	BT-Def+Expl+Ex
Framing	187	48	142	118
Epistemological	1	32	13	25
Demographic	3	1	3	11
None detected	10	120	43	47

Table 5 shows the bias-type distribution results from the secondary experiment described in Section 3.5, across the 201 human-labeled biased sentences. Under BT-Direct, 93% of bias is classified as framing — the most surface-visible type, suggesting the model defaults to pattern-matching on emotionally charged language without deeper policy guidance. BT-Def causes the model to abstain frequently (59.7% "None"), reflecting under-confidence when given definitions alone without explanatory context. BT-Def+Expl produces the most balanced distribution (framing 70.6%, epistemological 6.5%, demographic 1.5%), while BT-Def+Expl+Ex achieves the best demographic detection (5.5%) at the cost of slightly higher undetected cases (23.4%). The near-invisibility of epistemological and demographic bias under BT-Direct confirms that without detailed policy context, the model cannot distinguish subtle bias types from neutral writing. The demographic detection trend ($1 \rightarrow 3 \rightarrow 11$ across BT-Def through BT-Def+Expl+Ex) further shows that concrete examples are uniquely necessary for detecting the rarest bias type. These patterns suggest that detailed policy explanation is more valuable than examples for overall bias-type calibration, though

examples specifically improve detection of underrepresented types like demographic bias.

6 Community Language Analysis

While topic analysis identifies *what* a sentence is about, discourse community analysis reveals *how* it is written. To examine whether model errors are driven by topic content or by language style independently, we classified sentences into eight discourse communities (medical, scientific, political, cultural, business, technical, legal, and sports) based on keyword matching. Sentences with no domain-specific keywords were assigned to a general category. For example, a sentence containing words like "treatment" or "clinical" was assigned to the medical community, while a sentence containing "government" or "policy" was assigned to the political community. Sentences with no domain-specific keywords were assigned to the general community. AI-human disagreement rates were then computed per community.

Table 6: AI-Human Disagreement Rate by Discourse Community (n=400)

Discourse Community	N	Disagreement Rate
Medical	14	64.3%
Business	11	45.5%
Technical	9	44.4%
Scientific	16	43.8%
Sports	32	34.4%
Cultural	63	35.2%
Legal	9	31.4%
General	165	27.3%
Political	81	27.2%
Total	**400**	**31.0%**

Table 6 reveals that medical discourse exhibits by far the highest error rate (64.3%). Breaking this down, 28.6% of medical sentences were over-flagged as biased (false positives) and 35.7% represented missed bias (false negatives). These asymmetric errors suggest two failure modes: the model conflates authoritative clinical language with bias, while simultaneously failing to detect subtle value-laden framing in medical writing—for instance, expressions of clinical certainty or patient value judgments that violate NPOV.

Business (45.5%), technical (44.4%), and scientific (43.8%) discourse all show substantially elevated disagreement rates. In contrast, general (27.3%) and political (27.2%) content achieves the lowest disagreement, likely because political bias tends to be lexically explicit—partisan adjectives, unattributed opinions, loaded verbs—which aligns more closely with the surface patterns the model has internalized during pre-training.

These findings have direct implications for adaptive web deployments. A single general-purpose model with a general-purpose NPOV prompt is insufficient for reliable moderation across all professional discourse communities. Domain-specialized prompting, fine-tuning on domain-specific examples, or human-in-the-loop workflows are necessary for high-stakes domains such as medical and scientific content.

It is important to note that high-disagreement communities such as medical (N=14), business (N=11), technical (N=9), and legal (N=9) have small sample sizes, which limits the statistical reliability of these estimates. The general (N=165) and political (N=81) communities have substantially larger samples, making their lower disagreement rates more robust. These limitations are discussed further in Section 7.4.

7 Discussion

7.1 Prompt Strategy as a Core Design Variable

Our results confirm that prompt strategy is the single most consequential variable in LLM-based NPOV moderation. For `claude-sonnet-4-5`, the accuracy gap between the worst (P1: 51.75%) and best (P3: 69.0%) prompts exceeds 17 percentage points—larger than the gap between model generations. This finding underscores that in adaptive web deployments, prompt design is not an implementation detail but a core architectural decision that determines system reliability. Organizations deploying LLMs for content moderation must invest as much effort in prompt engineering as in model selection.

The relationship between few-shot examples and performance warrants careful interpretation. As confirmed by McNemar's tests reported in Section 4.1, the P3-to-P4 drop for `claude-sonnet-4-5` is not statistically significant, indicating that adding examples neither reliably improves nor hurts performance for capable models. Interestingly, few-shot examples do benefit `claude-3-haiku` significantly, suggesting that example-based guidance may compensate for limited instruction-following capacity in smaller models. This tradeoff is also visible at the bias-type level: BT-Def+Expl produces the most calibrated framing detection, while BT-Def+Expl+Ex improves demographic detection at the cost of increased undetected cases, suggesting that examples optimize for rare bias types at the expense of general accuracy.

7.2 Model Generation and Default Behavior

Newer, more capable models do not uniformly outperform older models across all prompt conditions. `claude-sonnet-4-5` dramatically underperforms on direct instruction (near-random performance) but surpasses all other models when given rich policy context. This pattern suggests that instruction-aligned models are more conservative by default—they avoid strong judgments without sufficient context—but are highly responsive to policy-grounded prompting. Deploying a highly capable model with an under-specified prompt may therefore perform *worse* than a smaller model, while the same model with a well-designed policy prompt substantially outperforms all alternatives. This has important practical implications for organizations choosing between model size and prompt sophistication.

7.3 Risks for the Adaptive Web

From an ABIS perspective, our findings reveal a fundamental tension between the promise and the peril of AI-fueled content moderation. On one hand, LLMs can serve as scalable first-pass filters that assist human editors. On the other hand, their systematic failures in medical, scientific, and technical domains could suppress

legitimate expert content or allow biased health and science information to persist unchecked. Given that Wikipedia content is embedded throughout the adaptive web—cited in search results, ingested into LLM training corpora, and referenced by recommendation systems—errors in NPOV enforcement propagate far beyond the platform. This amplification dynamic makes the reliability of automated moderation a societal concern, not merely a platform engineering challenge.

7.4 Limitations

This study evaluates 400 sentences, limiting statistical power for fine-grained topic and community analyses. We test only Claude model variants; results may differ for GPT-4, Llama, or other families. This study does not compare against supervised models previously evaluated on WNC, as our research question concerns zero-shot and few-shot deployability rather than supervised benchmark performance. The bias-type results cannot be directly compared to the bias detection results, as the two experiments differ in task design; the bias detection experiment performs binary classification on 400 sentences, while the bias-type experiment performs four-way classification on the 201 human-labeled biased sentences using a separate prompt set. Discourse community classification uses keyword matching, which is a coarse proxy. Future work should explore larger datasets, additional model families, domain-adapted fine-tuning, and longitudinal evaluation on live Wikipedia content. Additionally, future studies should explore chain-of-thought or explanation-eliciting prompts that require the model to justify its classification, providing interpretability into why sentences are flagged as biased or neutral.

8 Conclusion

We have presented a systematic, multi-dimensional evaluation of LLM-based NPOV bias detection on Wikipedia using the Wikimedia Neutrality Corpus. Across three Claude model generations and four prompt strategies, `claude-sonnet-4-5` with NPOV guidelines and detailed explanation achieves the best performance at 69.0% accuracy—a dramatic improvement over the near-random direct-instruction baseline (51.75%) for the same model, and outperforming all other model and prompt combinations evaluated.

Our topic-aware analysis reveals that Culture-domain content is handled most reliably, while Geography suffers from low recall. STEM topics show high precision but missed violations. Our secondary bias-type analysis shows that framing dominates across all prompt conditions, while demographic bias specifically requires concrete examples to surface reliably. Most strikingly, our discourse community analysis identifies medical text (64.3% disagreement), business (45.5%), and technical language (44.4%) as the highest-risk communities for automated moderation failures, driven by asymmetric false positive and false negative patterns that reflect distinct linguistic confounds in each domain.

These findings make two contributions to the ABIS 2026 research agenda. First, they provide concrete empirical evidence that LLM performance on adaptive web tasks is highly sensitive to both prompt design and discourse community—the same model can range from near-random to reliable depending on these choices. Second, they identify the systematic failure modes that practitioners must address before deploying AI-based NPOV enforcement at scale. As Wikipedia and the broader adaptive web increasingly rely on LLMs for content quality assurance, ensuring these systems perform reliably across all discourse communities—not just general and political text—is both a technical challenge and a matter of equitable information access.

Acknowledgments

The authors thank the Anthropic API team for providing access to the Claude models used in this study, and the Wikipedia editor community for their invaluable contributions to the Wikimedia Neutrality Corpus. We also thank the ABIS 2026 organizers for creating a forum to discuss the societal effects of AI-fueled personalization on the adaptive web.

References

[1] Ashik Ahamed, Max Wang, Amity Ramona Mentis-Cort, and Jeanna Matthews. 2026. An analysis of Wikipedia's special tags and their implications for the nuanced spectrum between human edits and bot edits. *International Conference on Virtual Learning* 21 (2026), 15–26. doi:10.58503/icvl-v21y202601

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[3] Christoph Hube and Besnik Fetahu. 2018. Detecting biased statements in Wikipedia. In *Companion Proceedings of The Web Conference 2018*. 1779–1786.

[4] Eli Pariser. 2011. *The Filter Bubble: What the Internet is Hiding from You.* Penguin Press. https://www.penguinrandomhouse.com/books/309214/the-filter-bubble-by-eli-pariser/

[5] Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2020. Automatically neutralizing subjective bias in text. In *Proceedings of the aaai conference on artificial intelligence*, Vol. 34. 480–489.

[6] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.

[7] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, Jonathan May, Ekaterina Shutova, Aurelie Herbelot, Xiaodan Zhu, Marianna Apidianaki, and Saif M. Mohammad (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, USA, 75–86. doi:10.18653/v1/S19-2010

A Prompt Templates

All four prompt templates used in this study are reproduced below verbatim. The placeholder `{sentence}` is replaced with the target sentence at inference time. All prompts instruct the model to return a single word (*biased* or *neutral*).

P1 — Direct Instruction

```
Classify the following sentence as biased or neutral.
Return only one word: biased or neutral.

Sentence: {sentence}
```

P2 — NPOV Guidelines

```
Act as an expert Wikipedia editor. Classify if
this sentence is neutral or biased according to
Wikipedia's Neutral Point of View (NPOV) policy.

Wikipedia NPOV Policy:
- Avoid subjective and judgmental language (use
disinterested, impartial tone)
- Avoid phrasing that implies how 'believable' a
```

```
statement is
- Avoid stereotypes about gender, race, or other
demographics
- Avoid presenting opinions as facts (opinions
should be attributed)

Even slight violations should be classified as
biased.

Sentence: {sentence}

Return only one word: biased or neutral.
```

P3 — NPOV Guidelines + Explanation

```
Act as an expert Wikipedia editor. Classify if
this sentence is neutral or biased according to
Wikipedia's Neutral Point of View (NPOV) policy.

Wikipedia NPOV Policy (4 Core Principles):

1. AVOID SUBJECTIVE AND JUDGMENTAL LANGUAGE
  - Use disinterested, impartial tone
  - Avoid language that sympathizes with or disparages
the subject
  - Remove emotionally charged adjectives and
value-laden words

2. AVOID PHRASING THAT IMPLIES HOW 'BELIEVABLE' A
STATEMENT IS
  - Do not manipulate perceived truth or certainty
  - Avoid words that boost or hedge believability
inappropriately
  - Present information neutrally without signaling
doubt or confidence

3. AVOID STEREOTYPES ABOUT GENDER, RACE, OR OTHER
DEMOGRAPHICS
  - No generalizations about demographic groups
  - No evaluative statements based on identity
categories
  - Treat individuals as individuals, not representatives
of groups

4. AVOID PRESENTING OPINIONS AS FACTS
  - Opinions must be attributed to a person or
group
  - Distinguish between verifiable facts and subjective
views
  - Use phrases like 'according to X' or 'critics
argue'

IMPORTANT: Even SLIGHT violations should be classified
as biased.

Sentence: {sentence}

Return only one word: biased or neutral.
```

P4 — NPOV Guidelines + Explanation + Examples

```
Act as an expert Wikipedia editor. Classify if
this sentence is neutral or biased according to
Wikipedia's Neutral Point of View (NPOV) policy.

Wikipedia NPOV Policy (4 Core Principles):

1. AVOID SUBJECTIVE AND JUDGMENTAL LANGUAGE
  - Use disinterested, impartial tone
  - Avoid language that sympathizes with or disparages
the subject
  - Remove emotionally charged adjectives and
value-laden words
  Examples of violations:
    × 'brilliant scientist' → ✓ 'scientist'
    × 'tragic incident' → ✓ 'incident'
    × 'unfortunately failed' → ✓ 'failed'
    × 'proudly announced' → ✓ 'announced'

2. AVOID PHRASING THAT IMPLIES HOW 'BELIEVABLE' A
STATEMENT IS
  - Do not manipulate perceived truth or certainty
  - Avoid words that boost or hedge believability
inappropriately
  - Present information neutrally without signaling
doubt or confidence
  Examples of violations:
    × 'proves that climate change is real' → ✓
'indicates climate change'
    × 'clearly demonstrates' → ✓ 'demonstrates'
or 'shows'
    × 'allegedly committed' → ✓ 'was accused
of'
    × 'claims to cure cancer' → ✓ 'is used to
treat cancer'

3. AVOID STEREOTYPES ABOUT GENDER, RACE, OR OTHER
DEMOGRAPHICS
  - No generalizations about demographic groups
  - No evaluative statements based on identity
categories
  - Treat individuals as individuals, not representatives
of groups
  Examples of violations:
    × 'Women are naturally more empathetic'
    × 'Asian students excel at mathematics'
    × 'The French are known for their romance'
    × 'Men tend to be more aggressive'

4. AVOID PRESENTING OPINIONS AS FACTS
  - Opinions must be attributed to a person or
group
  - Distinguish between verifiable facts and subjective
views
  - Use phrases like 'according to X' or 'critics
argue'
  Examples of violations:
    × 'This is the best approach' → ✓ 'Experts
consider this an effective approach'
    × 'The policy is harmful' → ✓ 'Critics argue
the policy is harmful'
    × 'Obama is a hypocrite' → ✓ 'Critics called
Obama a hypocrite'
    × 'The movie is terrible' → ✓ 'The movie
received negative reviews'
```

```
IMPORTANT: Even SLIGHT violations should be classified
as biased.

Sentence: {sentence}

Return only one word: biased or neutral.
```

B Bias-Type Detection Prompt Templates

All four bias-type detection prompt templates used in the secondary experiment are reproduced below verbatim. The placeholder {sentence} is replaced with the target sentence at inference time. All prompts instruct the model to return a single word (*framing*, *epistemological*, *demographic*, or *none*). These prompts are distinct from the P1–P4 neutrality classification prompts in Appendix A.

BT-Direct — Direct Instruction

```
You are analyzing bias in a Wikipedia sentence
according to Wikipedia's Neutral Point of View
(NPOV) policy.

There are only four possible bias types: framing,
epistemological, demographic and none.

IMPORTANT:
Return ONLY ONE WORD: framing, epistemological,
demographic, or none.
Do NOT provide explanation. Do NOT add punctuation.

Sentence: {sentence}

Answer:
```

BT-Def — Definition Only

```
You are analyzing bias in a Wikipedia sentence.

There are only four possible labels:

1. framing — Emotionally charged or value-laden
wording
2. epistemological — Language expressing unwarranted
certainty
3. demographic — Generalizations about demographic
groups
4. none — No clear bias present

Return ONLY ONE WORD: framing, epistemological,
demographic, or none.

Sentence: {sentence}
```

BT-Def+Expl — Definition + Explanation

```
You are analyzing bias in a Wikipedia sentence
according to Wikipedia's Neutral Point of View
(NPOV) policy.

There are only four possible bias types:

1. framing
  Definition: Use of subjective and judgmental
language that reveals the author's attitude
  Explanation: Word choice that sympathizes with
or disparages the subject, emotionally charged
terms, value-laden words

2. epistemological
  Definition: Language that implies unwarranted
certainty or manipulates how believable a statement
is
  Explanation: Presenting opinions as facts, stating
contested assertions without attribution, hedging
or boosting believability

3. demographic
  Definition: Language containing stereotypes or
generalizations about demographic groups
  Explanation: Evaluative statements about gender,
race, ethnicity, nationality, religion, or other
demographic categories

4. none
  Definition: No clear bias present
  Explanation: The sentence uses neutral, factual
language without subjective judgment, unwarranted
certainty, or demographic generalizations

IMPORTANT:
Return ONLY ONE WORD: framing, epistemological,
demographic, or none.
Do NOT provide explanation. Do NOT add punctuation.

Sentence: {sentence}

Answer:
```

BT-Def+Expl+Ex — Definition + Explanation + Examples

```
You are analyzing bias in a Wikipedia sentence
according to Wikipedia's Neutral Point of View
(NPOV) policy.

There are only four possible bias types:

1. framing
  Definition: Use of subjective and judgmental
language that reveals the author's attitude
  Explanation: Word choice that sympathizes with
or disparages the subject, emotionally charged
terms, value-laden words
  Examples: 'brilliant,' 'tragic,' 'unfortunately,'
'clearly,' 'obviously'

2. epistemological
  Definition: Language that implies unwarranted
certainty or manipulates how believable a statement
is
  Explanation: Presenting opinions as facts, stating
contested assertions without attribution, hedging
or boosting believability
  Examples: 'proves that,' 'demonstrates,' 'obviously
true,' 'allegedly,' 'claims to'

3. demographic
  Definition: Language containing stereotypes or
```

```
generalizations about demographic groups
  Explanation: Evaluative statements about gender,
race, ethnicity, nationality, religion, or other
demographic categories
  Examples: 'Women are naturally more empathetic,'
'the Chinese tend to,' implicit stereotypes

4. none
  Definition: No clear bias present
  Explanation: The sentence uses neutral, factual
language without subjective judgment, unwarranted
certainty, or demographic generalizations

IMPORTANT:
Return ONLY ONE WORD: framing, epistemological,
demographic, or none.
Do NOT provide explanation. Do NOT add punctuation.

Sentence: {sentence}

Answer:
```