Large Language Models as Reader for Bias Detection

Anonymous ACL submission

Abstract

Detecting bias in media content is crucial for maintaining information integrity and promoting inclusivity. Traditional methods analyze text from the writer's perspective, which analyzes textual features directly from the writer's intent, leaving the reader's perspective underexplored. This paper investigates whether Large Language Models (LLMs) can be leveraged as readers for bias detection by generating reader-perspective comments. Experiments are conducted on the BASIL (news bias) and BeyondGender (gender bias) datasets with LLMs Gemma-7B, Phi-3-3.8B, Llama3.1-8B, Llama3.1-70B, and GPT4. The results demonstrate the effectiveness of reader-perspective comments for open-source LLMs, achieving performance comparable to GPT4's. The findings highlight the significance of emotion-related comments, which are generally more beneficial than value-related ones in bias detection. In addition, experiments on Llamas show that comment selection ensures consistent performance regardless of model sizes and comment combinations. This study is particularly beneficial for small-size open-source LLMs.

1 Introduction

The rapid expansion of digital media has intensified concerns regarding biased content, characterized by deviations from objective representation that favor particular viewpoints, groups, or outcomes, whether introduced intentionally or unintentionally. Identifying such biased language (Bias Detection) in media content, such as news articles and social media posts, has become a critical challenge (Garg et al., 2023; Rodrigo-Ginés et al., 2024). Traditional methods adopt writer's perspective to analyze textual features directly tied to the author's intent. They assume that bias originates from the writer's language and framing, implicitly adopting Hall's "encoding" perspective (Hall, 2019).[1] However, existing methods neglect the "decoding" process where readers actively interpret and construct meaning (Rosenblatt, 1969).

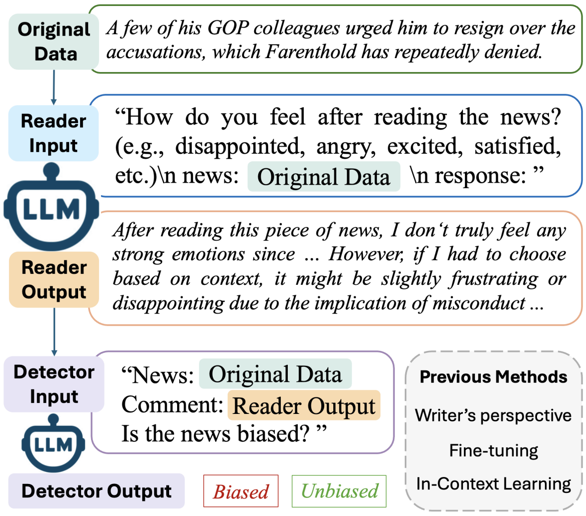

Figure 1: The experiment workflow. Step-1: The Reader generates comments based on the original data upon receiving it. Step-2: The generated comment is appended to the original data, creating a new input. Step-3: The Detector LLM makes an inference using the concatenated input.

Large Language Models (LLMs) have shown remarkable capabilities in text understanding and generation (Yang et al., 2024), often being used for data synthesis and reasoning explanation. While they have been used for bias detection, their potential as a **Reader**, observing data and generating rational or emotional comments instead of from the writer's perspective, remains underexplored. On the other hand, bias detection datasets are annotated by human audience (readers) rather than the content producer (writer). Therefore, we intuitively utilize LLMs to align with human annotation

[1]Hall's communication model posits that meaning emerges through the encoding by producers with various signs and decoding by viewers with their own framework of knowledge.

through this process, shifting the lens from what the writer says to how readers perceive them. In addition, inspired by the fact that human perceptions of bias can be influenced by user comments (Houston et al., 2011; Lee, 2012; Gearhart et al., 2020), we wonder whether LLMs can be leveraged as peer-readers to simulate this dynamic and enhance bias detection capability.

The Research Questions are as follows: **RQ1**) Are reader-perspective comments effective in bias detection? **RQ2**) Can LLMs influenced by peer-reader's comments? **RQ3**) What kind of comment generation policies are most / more beneficial? **RQ4**) Whether a trained selector deciding to add the comment or not improves the performance?

Experiments are designed as follows: Initially, we utilize an LLM to generate comments that capture diverse viewpoints or express emotions evoked by the content. Then, LLMs make the inference with these comments combined with the original content. We evaluate on the news bias dataset BASIL (Fan et al., 2019) and the gender bias dataset BeyondGender (Luo et al., 2025) with different LLMs backbones: GPT, Phi, Gemma, and Llama, primarily focusing on small-size LLMs.

The main contributions and findings are as follows: **1)** A novel perspective utilizing LLMs as Reader to generate comments for bias detection, which is effective on news bias and gender bias detection (RQ1), **2)** Findings that small-size LLMs' performance is significantly improved by the influence of peer-reader's comments (RQ2), resource-efficient for computing. **3)** Findings that emotion-related comments are generally more beneficial than value-related ones and that comments vary with the reader's gender (RQ3), providing insights into how to utilize comments effectively in biased content analysis. **4)** Findings that comment selection may be helpful, yet the positive effects depend on the backbone, requiring further analysis. (RQ4).

2 Experiment Design

The workflow is illustrated in Figure 1. By default, we employ a greedy strategy, where the best policy comments are appended to the original data (Section 4.1 and 4.2).[2]

Further we explore the comment selection setting (Fig. 3 in Appx. C), where a Selector evaluates the usefulness of each comment and decides whether it should be appended with. (Section 4.3).

2.1 Reader-Perspective Design

We categorize reader-perspective comments into two primary dimensions: General and Individual.

General Perspective. Motivated by Stratton (2021), this dimension examines the external and rational aspects of content, focusing on:

1) **Portrayals** of target parties or groups: Selective emphasis on certain parties can influence public perception. Assessing how specific parties or groups are depicted in the content helps identify potential biases in media coverage.

2) **Values**: Media / User outlets may unintentionally or intentionally reflect certain values, influencing audience interpretation. Analyzing the values expressed in the content reveals whether they align with particular political ideologies.

Individual Perspective. Motivated by Han and Arpan (2017), this dimension explores the internal and emotional responses elicited by content:

1) **Emotions**: Identifying the emotions evoked by the content—such as anger, sadness, or joy—can indicate the presence of bias.

2) **Sharing Willingness**: Assessing the likelihood of readers sharing the content. A higher inclination to share may suggest that the content resonates or conflicts with the reader's emotions / beliefs, potentially indicating bias in the reporting.

3) **Life Impact**: Content perceived as impactful on life may be more engaging or persuasive, which can be influenced by the way it is presented.

2.2 Component Design

Reader: Comment Generation. We employ Llama3.1-70B (Grattafiori et al., 2024) to produce reader-perspective comments from both dimensions. For each sample, we instruct the LLM with "If yes, please specify" under the policies, as shown in the Appendix B Table 4 and 5.[3]

Detector: Bias Detection. Provided with the original data and selected positive comments, the Detector (an LLM) is instructed to detect bias in a zero-shot setting. The prompt is, " $news$: + original_data + $comment$: + generated_comments + $Is\ the\ news\ biased?$ ". The word "news" is replaced with an appropriate term based on the data.

Selector: Comment Selection. We utilize generated comments to train a comment selector

[2] The best policy is observed by the training set, which leads to the best results.

[3] Preliminary experiments show that combining a large-size LLM with simple prompts yields better comments. Simple prompts also lead to better performance during inference.

LLM	BASIL Inf/ Lex / non		BeyondGender Sexism		Gender		Misogyny		Misandry	
	micro-F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC
GPT4	**0.83**	**0.89**	0.84	0.74	**0.51**	**0.67**	0.81	0.71	**0.25**	0.42
GPT4 + AUG	0.82	0.88	**0.85** ↑	**0.75** ↑	0.50	0.66	**0.82** ↑	**0.73** ↑	0.21	**0.52** ↑
Existing SOTA	-	0.81	0.79	0.67	0.40	0.30	0.69	0.59	0.19	0.30
Llama-70B	0.54	0.54	0.22	0.30	0.40	0.28	0.12	0.29	0.09	**0.89**
Llama-70B + AUG	0.64 ↑	0.76	0.83 ↑	0.72 ↑	0.39	0.26	**0.83** ↑	**0.73** ↑	0.16	0.20
Llama-8B	0.62	0.62	0.73	0.61	0.32	0.33	0.72	0.62	0.16	0.43
Llama-8B + AUG	0.70 ↑	0.80	0.80 ↑	0.70 ↑	0.41 ↑	**0.46** ↑	0.81 ↑	0.71 ↑	0.18	0.34
Phi-3-3.8B	0.28	0.52	0.83	0.72	0.33	0.22	0.78	0.69	0.14	0.47
Phi-3-3.8B + AUG	0.73 ↑	0.83 ↑	**0.84** ↑	**0.73** ↑	**0.42** ↑	0.35 ↑	0.80 ↑	0.70 ↑	0.20 ↑	0.60 ↑
Gemma-7B	0.27	0.51	0.51	0.47	0.32	0.22	0.55	0.51	0.19	0.73
Gemma-7B + AUG	**0.80** ↑	**0.87** ↑	0.73	0.61	0.40	0.32 ↑	0.76 ↑	0.64 ↑	**0.21** ↑	0.73

Table 1: Main results of baselines and comment-augmented models (+ AUG). The values are F1-scores and accuracy. The best results among open-source models and closed-source GPT4 are bolded separately. ↑ denotes that +AUG surpasses **both** the baseline and existing SOTA (Maab et al. (2023) for BASIL and Luo et al. (2025) for BeyondGender). The McNemar's test between baselines and comment-augmented models (+AUG), $p < 0.05$.

(BERT Devlin et al., 2019) capable of distinguishing between positive (useful) and negative (unuseful) comments. Training details are in Appx. C.

3 Experiment Settings

3.1 Datasets

Our method is evaluated on the following datasets:

BASIL (Fan et al., 2019). It is a news bias detection dataset, with around 8K sentences labeled as informational bias, lexical bias, or unbiased. Following the formulation of the dataset, we classify the news data as "Inf", "Lex", or "non-bias". [4]

BeyondGender (Luo et al., 2025). It is a gender bias detection dataset, with over 13K English posts collected from social media. Following Luo et al.'s settings, we separately detect the 4 bias-related labels: sexism, gender, misogyny, and misandry.

The statistics of original datasets are in Table 2.

Dataset	Label	Train	Test
BASIL	Inf	349	123
	Lex	138	32
	Non-bias	2,067	641
BeyondGender	Sexism	4,381	485
	Gender	5,233	367
	Misogyny	5,233	367
	Misandry	5,233	367

Table 2: Statistic of the original datasets.

3.2 Models and SOTAs

We evaluate the detection performance of Phi-3-3.8B (Abdin et al., 2024), Gemma-7B (Team et al., 2024), Llama3.1-8B, Llama3.1-70B (Grattafiori et al., 2024), and GPT4 (OpenAI, 2023). LLMs are utilized with their default hyperparameters.

For BASIL, the state-of-the-art (SOTA) method for three-class classification is proposed by Maab et al. (2023), which utilizes supervised learning with augmented training data. For BeyondGender, the SOTA is Llama's few-shot in-context learning performance reported in Luo et al. (2025).

4 Results

4.1 Main Results

The main results with greedy strategy, in Table 1, address **RQ1** (effectiveness) and **RQ2** (peer-reader's comments).

The effectiveness of our method is evidenced by substantial and consistent improvements in both F1-score and accuracy (mean of three runs) achieved by Llama, Phi, and Gemma, comparing the baselines and +AUG. Regarding model size, while baseline Llama-70B performs much worse than Llama-8B, they achieve comparable results with comment augmentation (Llama-70B+AUG vs. Llama-8B+AUG), underscoring the effectiveness of reader-perspective comments. (RQ1)

Even though the results baselines vary, the generated comments consistently improve small-size open-source models' performance. In contrast, comments provide limited benefit for GPT-4, whose high performance is likely attributed to its extensive pre-training on sensitive topics with a vast volume of labeled data. Notably, small-size models perform on par with GPT-4 on both datasets, indicating that small-size LLMs are more susceptible to peer-reader(LLama-70B)'s comments compared to LLama-70B and GPT4. (RQ2)

[4]According to Maab et al. (2023), prior work utilizing BASIL with inconsistencies in the task formulation, which are derived from how these labels are interpreted and used.

LLM	BeyondGender							
	Sexism		Gender		Misogyny		Misandry	
	F1	ACC	F1	ACC	F1	ACC	F1	ACC
Best of open-source in Table 1	0.84	0.73	**0.42**	**0.46**	0.83	0.73	0.21	0.89
Existing SOTA	0.79	0.67	0.40	0.30	0.69	0.59	0.19	0.30
Llama-8B	0.73	0.61	0.32	0.33	0.72	0.62	0.16	0.43
Top-1 (Greedy Strategy)	0.80	0.70	0.41	**0.46**	0.81	0.71	0.18	0.34
Top-1 + Selector	0.84 ↑	0.74 ↑	0.40	0.37	0.84 ↑	0.75 ↑	0.17	0.26
Top-2	0.75	0.62	0.40	0.43	0.76	0.65	0.12	0.41
Top-2 + Selector	0.84 ↑	0.73 ↑	0.39	0.40	0.83 ↑	0.72 ↑	0.17 ↑	0.22
Random-1	0.72	0.64	0.40	0.42	0.78	0.66	0.13	0.40
Random-1 + Selector	0.83 ↑	0.73 ↑	**0.42** ↑	0.40	0.84 ↑	0.75 ↑	0.18 ↑	0.24
Random-2	0.73	0.61	0.35	0.43	0.72	0.61	0.15	0.45
Random-2 + Selector	**0.85** ↑	**0.75** ↑	0.40 ↑	0.38	**0.85** ↑	**0.76** ↑	0.18 ↑	0.23

Table 3: Results of different combinations of comments using Llama-8B as Detector. Top-k/Random-k: choose comments from the top/random k policies, whether positive or negative, and provide them together to Detector. Top-k/Random-k + selector: after choosing the top-k/random-k comments, only provide the positive comment(s) together to detector. ↑ denotes the improvement of +Selector. Best results are in bold.

4.2 Policy Analysis

To answer **RQ3**, we juxtapose the results of all policies (Table 4 and 5) in Fig. 2 for Llama-8B. [5]

For BASIL, individual perspectives are generally above the average and the best policy is No.13. Surprisingly, the value-related or politial-party related comments (except for No.4 focusing on language) have negative impact on news bias detection.

For BeyondGender, each label achieves the best performance with policy No.11, 5, 10, and 10, respectively. Moreover, Sexism, Misogyny, and Miandry have a similar trend, with policies No.6-7 and 9-11 above the average. Specifically, the gender difference between policies No.7 vs 10 and 12 vs 13 leads to performance gaps, revealing the disparity of comments regarding the reader's gender.

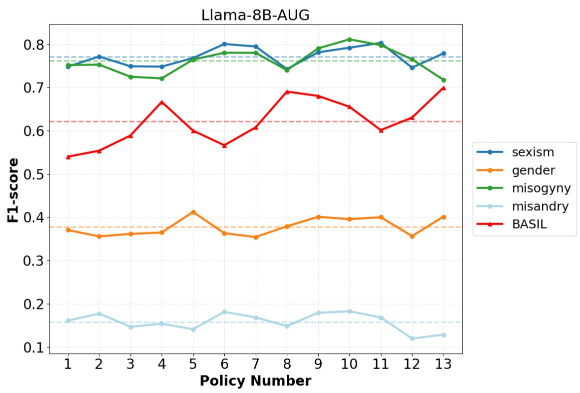

Figure 2: The F1-scores of each policy by Llama-8B. The red line with triangles is BASIL; the blue, orange, green, and light blue lines with circles are Sexism, Gender, Misogyny, and Misandry, respectively. The dashed lines indicate the averages. Policy No.1-6 are general perspectives and No.7-13 are individual perspectives.

[5]Figure 5 and 4 for Gemma and Phi-3 are in the Appx. D.

4.3 Selector Analysis

To address **RQ4**, we compare comment selection with greedy strategy and try several comment combinations, as detailed in Table 3. Compared to the Llama-8B baseline, both Top combinations significantly enhance performance, whereas both Random combinations offer little improvement. When comparing Top-1 to -2 and Random-1 to -2, it is evident that an increased number of comments can negatively impact performance, potentially due to the extended length. More results in Appx. E.

Although only the Top-1 policy surpasses the existing SOTA across all labels, the selector boosts performance to a comparable level regardless of the comment combinations. They suggest that the potential bottleneck of the Reader-Selector-Detector pipeline may be the quality of the comments and the accuracy of the selector. However, selectors do not work well with Gemma and Phi backbones (see Appendix E). Comment selection enhances the performance less than the greedy strategy. These findings provide a partially confirmed answer to RQ4, depending on the LLM backbones.

5 Conclusion

In this work, we explore leveraging LLMs as readers to generate reader-perspective comments for bias detection. Through the design of comment generation policies and experiments on various LLMs, the results demonstrate significant effectiveness and robustness in detecting bias for both news and gender bias. The findings highlight the potential of utilizing large-size LLMs as dynamic readers in various roles and small-size LLMs as efficient detectors for other types of content analysis.

Limitations

The experimental results suggest that a key bottleneck may lie in the quality of the generated comments, as LLama's performance stabilizes after comment selection. This indicates that the power of our method is closely tied to the quality of the generated comments. However, there is a lack of standardized methods for evaluating the upper-bound of generation quality across different Large Language Models. A potential avenue for future improvement could involve developing self-improvement strategies to enhance comment quality.

Additionally, although our findings highlight the significance of emotion-related comments in bias detection, the exact nature of this relationship remains unclear and warrants further investigation. We also observe that comments are particularly beneficial when the baseline performance is suboptimal. In contrast, for large closed-source models like GPT-4, which already exhibit strong bias detection capabilities, the impact of comment augmentation is less pronounced.

Since our focus is the small-size open-source LLMs, few large-size and closed-source models are evaluated.

Ethical Considerations

it is crucial to acknowledge the ethical implications and potential risks associated with the use of Large Language Models (LLMs). LLMs are trained on vast datasets that may contain inherent biases, which can lead to the generation of content that reflects and potentially amplifies these biases. Despite the straightforwardness and effectiveness of our method, the generated comments are not actively monitored, raising concerns about fairness and the potential amplification of existing societal biases, including gender and political biases. The other issue is the risk of contaminating online data if these comments are released or distributed.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Angana Borah and Rada Mihalcea. 2024. Towards implicit bias detection and mitigation in multi-agent LLM interactions. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9306–9326, Miami, Florida, USA. Association for Computational Linguistics.

Patricia Chiril, Farah Benamara, and Véronique Moriceau. 2021. "be nice to your wife! the restaurants are closed": Can gender stereotype detection improve sexism classification? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2833–2844, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Lisa Fan, Marshall White, Eva Sharma, Ruisi Su, Prafulla Kumar Choubey, Ruihong Huang, and Lu Wang. 2019. In plain sight: Media bias through the lens of factual reporting. 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)*, pages 6343–6349, Hong Kong, China. Association for Computational Linguistics.

Tanmay Garg, Sarah Masud, Tharun Suresh, and Tanmoy Chakraborty. 2023. Handling bias in toxic speech detection: A survey. *ACM Comput. Surv.*, 55(13s).

Sherice Gearhart, Alexander Moe, and Bingbing Zhang. 2020. Hostile media bias on social media: Testing the effect of user comments on perceptions of news bias and credibility. *Human behavior and emerging technologies*, 2(2):140–148.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Stuart Hall. 2019. Encoding—decoding (1980). In *Crime and media*, pages 44–55. Routledge.

Yi-Hsing Han and Laura Arpan. 2017. The effects of news bias-induced anger, anxiety, and issue novelty on subsequent news preferences. *Advances in Journalism and Communication*, 5(4):256–277.

J Brian Houston, Glenn J Hansen, and Gwendelyn S Nisbett. 2011. Influence of user comments on perceptions of media bias and third-person effect in online news. *Electronic News*, 5(2):79–92.

Eun-Ju Lee. 2012. That's not the way it is: How user-generated comments on the news affect perceived media bias. *J. Comp.-Med. Commun.*, 18(1):32–45.

Nayeon Lee, Belinda Z. Li, Sinong Wang, Pascale Fung, Hao Ma, Wen-tau Yih, and Madian Khabsa. 2021. On unifying misinformation detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5479–5485, Online. Association for Computational Linguistics.

Yuanyuan Lei, Ruihong Huang, Lu Wang, and Nick Beauchamp. 2022. Sentence-level media bias analysis informed by discourse structures. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10040–10050, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xuan Luo, Li Yang, Han Zhang, Geng Tu, Qianlong Wang, Keyang Ding, Chuang Fan, Jing Li, and Ruifeng Xu. 2025. Beyondgender: A multifaceted bilingual dataset for practical sexism detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24750–24758.

Iffat Maab, Edison Marrese-Taylor, and Yutaka Matsuo. 2023. An effective approach for informational and lexical bias detection. In *Proceedings of the Sixth Fact Extraction and VERification Workshop (FEVER)*, pages 66–77, Dubrovnik, Croatia. Association for Computational Linguistics.

Iffat Maab, Edison Marrese-Taylor, Sebastian Padó, and Yutaka Matsuo. 2024. Media bias detection across families of language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4083–4098, Mexico City, Mexico. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774. Accessed: 2025-01-11.

Francisco-Javier Rodrigo-Ginés, Jorge Carrillo de Albornoz, and Laura Plaza. 2024. A systematic review on media bias detection: What is media bias, how it is expressed, and how to detect it. *Expert Systems with Applications*, 237:121641.

Louise M Rosenblatt. 1969. Towards a transactional theory of reading. *Journal of Reading Behavior*, 1(1):31–49.

Indira Sen, Dennis Assenmacher, Mattia Samory, Isabelle Augenstein, Wil Aalst, and Claudia Wagner. 2023. People make better edits: Measuring the efficacy of LLM-generated counterfactually augmented data for harmful language detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10480–10504, Singapore. Association for Computational Linguistics.

Connor Stratton. 2021. *Bias in Reporting on Politics*. Focus on Media Bias. North Star Editions.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Esther van den Berg and Katja Markert. 2020. Context in informational bias detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6315–6326, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. 2024. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Trans. Knowl. Discov. Data*, 18(6).

A Related Work

The study of bias and discrimination has deep roots in psychology and social science. Research in these fields has shown that human perception of bias is often influenced by cognitive frameworks, social norms, and individual perspectives.

Traditional methods for bias detection often rely on supervised learning, focusing on identifying the appropriate contextual information for training (van den Berg and Markert, 2020; Lee et al., 2021; Lei et al., 2022) and training data augmentation through rule-based alterations or translation (Chiril et al., 2021; Maab et al., 2023). Recent advancements in Large Language Models (LLMs) have simplified data augmentation (Sen et al., 2023) and also bring new possibilities for bias detection (Yang et al., 2024). For instance, Maab et al. (2024) explore the potential of LLMs in news bias detection using prompt-based techniques while Borah and Mihalcea (2024) leverage multi-agent LLM interactions to detect gender bias.

However, existing studies primarily analyze text from the writer's perspective. On the other hand, research in psychology and social science has discovered the importance of external perspectives in bias perception (Houston et al., 2011; Lee, 2012; Gearhart et al., 2020). Drawing inspiration from this, we utilize LLMs as readers to generate reader-perspective comments, providing additional signals for bias detection.

B Comment Generation Prompt

Table 4 and 5 are the prompts for generating reader-perspective comments for BASIL (news bias detection) and BeyondGender (gender bias detection), respectively.

C Training of Selector

The followings are the role, training, and labeling procedure.

Role & Workflow. Before Step 2 in the Figure 1, these reader-perspective comments are filtered by a fine-tuned model, such as BERT, to determine whether append it to the original data or not. This Reader-Selector-Detector workflow is illustrated in Figure 3.

Labeling. The labeling process are as follows: Initially, we record the LLM's prediction for each original sample. Then, we append the generated comment to the original data and observe LLM's prediction on the comment-augmented input. A

Figure 3: The workflow of the selector setting. Three roles: **Reader** for reader-perspective comments generation, **Selector** for positive (helpful) comment selection, and **Detector** for bias detection utilizing original data and positive comments combined.

comment is labeled as positive if it changes an incorrect prediction to correct, and negative if it alters a correct prediction to incorrect. With these comment-augmented inputs and labels, we train a selector (binary classifier).

Training. The hyperparameters are pre-defined. Specifically, the seed=42, learning rate=1e-5, optimizer=AdamW, and epoch=1 to 15. The train set of the original datasets are randomly divided to 70% for training and 30% for evaluation (dev set) in each epoch. The best selector (BERT) models are selected with the 30% dev set and used for comment selection during testing. The classification performance is between 80% to 92%. Only the test set of the original datasets is used for pipeline evaluation.

D Policy Analysis Figure

Figure 4 and 5 show the F1-scores of each policy by Phi-3-3.8B and Gemma-7B, respectively. Phi-3 has similar pattern to Llama while Gemma has less fluctuation among different policies.

E Selector Analysis Table

Table 6 shows results of Llama-3.1-8B with more combinations and a negative comment. The degraded performance of negative comments demonstrate the selection capability of Selector.

Table 7 shows the results of Phi-3-3.8B and Gemma-7B with comment selection.

No.	Prompt
General: Does the news depict a particular party's image?	
1	Does the news have a positive or negative impact on a particular party or politician?
2	Does the news highlight any achievements, failures, or controversies associated with a particular party or politician?
3	Are there any specific statements, quotes, or actions mentioned in the news that could be seen as promoting or criticizing a particular party or politician?
4	Does the news use biased or critical language to describe a particular party or politician:
General: What values are expressed, and which party do they align with:	
5	Do the values or policies mentioned in the news align more with a particular party's stance?
6	Are there any specific words or phrases used in the news that emphasize a particular party's values or policies?
Individual: What emotional response does the news elicit:	
7	How do you feel after reading the news? (e.g. angry, disappointed, excited, satisfied, etc.)
8	Does the news have a positive, negative, or neutral tone?
9	Does the news use sensational/exaggerated language to evoke a specific emotional response?
Individual: Do you want to share the news with others:	
10	Do you want to share the news with friends or family?
11	Is the news intriguing enough to spark discussion or sharing?
Individual: Does the news have an impact on your life:	
12	Is the news relevant to your personal life or interests?
13	Does the news mention specific policies or actions that could affect your life?

Table 4: Reader-perspective comment generation policies for news bias detection.

No.	Prompt
General: Does the post depict a particular gender's image:	
1	Does the post have a positive or negative impact on a particular gender?
2	Does the post highlight any achievements, failures, or controversies associated with a particular gender?
3	Does the post express dominance or objectification over a particular gender?
4	Does the post use biased or critical language to describe a particular gender?
General: What values are expressed, and which gender do they favor:	
5	Do the values mentioned in the post favor or advantage a particular gender? If yes, please specify the values and the gender.
6	Are there any specific words or phrases used in the post that emphasize values that advantage or disadvantage a particular gender?
Individual: What emotional response does the post elicit:	
7/10	How do you feel after reading the post if you are a man / woman? (e.g. angry, disappointed, excited, etc.)
8	Does the post have a positive, negative, or neutral tone?
9	Does the post use sensational or exaggerated language to evoke a specific emotional response?
Individual: Do you want to share the post with others:	
11	Is the post intriguing enough to spark discussion on sexism?
Individual: Does the post have an impact on your life:	
12	Does the post mention specific policies or actions that could affect you if you are a woman?
13	Does the post mention specific policies or actions that could affect you if you are a man?

Table 5: Reader-perspective comment generation policies for gender bias detection.

LLM	BeyondGender							
	Sexism		Gender		Misogyny		Misandry	
	F1	ACC	F1	ACC	F1	ACC	F1	ACC
Best of open-source in Table 1	0.84	0.73	**0.42**	**0.46**	0.83	0.73	0.21	0.89
Existing SOTA	0.79	0.67	0.40	0.30	0.69	0.59	0.19	0.30
Llama-8B	0.73	0.61	0.32	0.33	0.72	0.62	0.16	0.43
Top-1 (Greedy Strategy)	0.80	0.70	0.41	**0.46**	0.81	0.71	0.18	0.34
Top-1 + Selector	0.84	0.74	0.40	0.37	0.84	0.75	0.17	0.26
Top-2	0.75	0.62	0.40	0.43	0.76	0.65	0.12	0.41
Top-2 + Selector	0.84	0.73	0.39	0.40	0.83	0.72	0.17	0.22
Random-1	0.72	0.64	0.40	0.42	0.78	0.66	0.13	0.40
Random-1 + Selector	0.83	0.73	**0.42**	0.40	0.84	0.75	0.18	0.24
Random-2	0.73	0.61	0.35	0.43	0.72	0.61	0.15	0.45
Random-2 + Selector	**0.85**	**0.75**	0.40	0.38	**0.85**	**0.76**	0.18	0.23
Top-3	0.75	0.63	0.37	0.45	0.77	0.65	0.18	0.43
Top-3+ Selector	0.83	0.72	0.41	0.41	**0.85**	**0.76**	0.19	0.25
Random-3	0.72	0.60	0.38	0.43	0.74	0.62	0.14	0.48
Random-3 + Selector	0.84	0.74	0.41	0.38	0.84	0.75	0.18	0.25
Llama8B + negativeAUG	0.55	0.48	0.32	0.37	0.43	0.44	0.16	0.50

Table 6: More Results of different combinations of comments using Llama-8B as Detector. Top-k/Random-k: choose comments from the top/random k policies, whether positive or negative, and provide them together to Detector. Top-k/Random-k + selector: after choosing the top-k/random-k comments, only provide the positive comment(s) together to detector. +negativeAUG refers to appending one negative comment.

LLM	BASIL		BeyondGender							
	Inf/ Lex / non		Sexism		Gender		Misogyny		Misandry	
	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC
Phi-3-3.8B	0.28	0.52	0.83	0.72	0.33	0.22	0.78	0.69	0.14	0.47
Phi-3-3.8B + Selector	0.37	0.58	0.83	0.72	0.34	0.38	0.78	0.68	0.12	0.60
Gemma-7B	0.27	0.51	0.51	0.47	0.32	0.22	0.55	0.51	0.19	0.73
Gemma-7B + Selector	0.43	0.62	0.50	0.48	0.34	0.28	0.52	0.48	0.15	0.60

Table 7: Results of Phi-3-3.8B and Gemma-7B with comment selection.

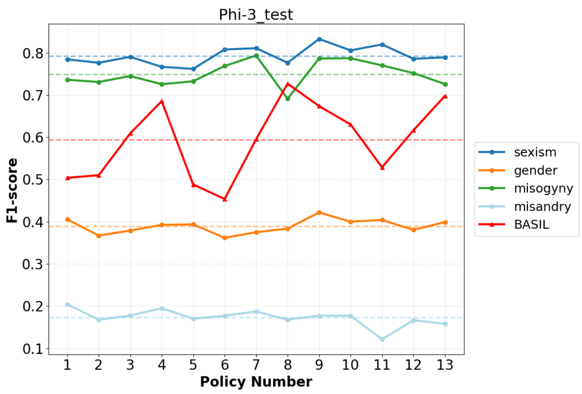

Figure 4: The F1-scores of each policy by Phi-3-3.8B. The red line with triangles is BASIL; the blue, orange, green, and light blue lines with circles are Sexism, Gender, Misogyny, and Misandry, respectively. The dashed lines indicate the averages. Policy No.1-6 are general perspectives and No.7-13 are individual perspectives.

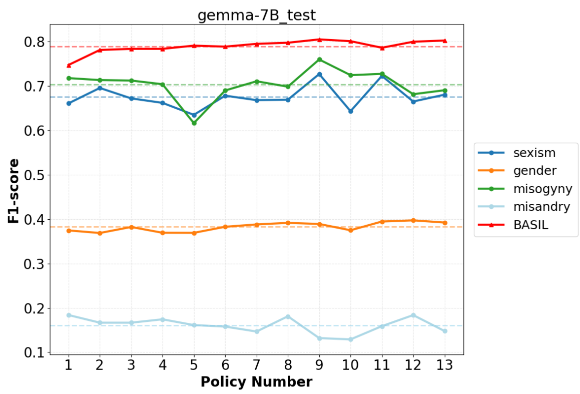

Figure 5: The F1-scores of each policy by Gemma-7B. The red line with triangles is BASIL; the blue, orange, green, and light blue lines with circles are Sexism, Gender, Misogyny, and Misandry, respectively. The dashed lines indicate the averages. Policy No.1-6 are general perspectives and No.7-13 are individual perspectives.