Large Language Models as Readers for Bias Detection

Anonymous ACL submission

Abstract

Detecting bias in media content is essential for ensuring information integrity and promoting inclusivity. Traditional methods often analyze text from the writer's perspective, leaving the reader's perspective underexplored. This paper introduces an innovative approach that leverages Large Language Models (LLMs) as readers to generate reader-perspective comments for bias detection. The most beneficial comments are selected by a selector and then utilized by an LLM to detect bias in the original data. We conduct experiments on the BASIL (news bias) and BeyondGender (gender bias) datasets with Llama3.1-8B. The results reveal the effectiveness of our method, achieving comparable performance to GPT4's. The findings highlight the significance of emotion-related comments, which are generally more beneficial than value-related ones in bias detection. Moreover, the reader's gender may influence comment quality. In addition, comment selection ensures consistent performance regardless of model sizes and comment combinations, demonstrating robustness and reliability.

1 Introduction

In the information era, identifying bias and discriminatory language (Bias Detection) in media content, such as news article and social media posts, has become a critical challenge (Garg et al., 2023; Rodrigo-Ginés et al., 2024). Large Language Models (LLMs) have shown remarkable capabilities in text understanding and generation (Yang et al., 2024), often being used for data synthesis and explaining reasoning. However, their potential as a **Reader**, observing data and generating rational or emotional comments instead of from the writer's perspective, remains underexplored.

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 intuitively leverage LLMs as Readers to simulate this dynamic and enhance bias detection capability. Initially, LLMs generate comments that capture diverse viewpoints or express emotions evoked by the content. These reader-perspective comments are then filtered by a fine-tuned model, such as BERT, to select the most beneficial ones. The selected comments provide additional contextual signals, enabling the LLMs to identify biases in the original text effectively. This Reader-Selector-Detector framework is illustrated in Figure 1.

Figure 1: The framework of the proposed method. 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.

The Research Questions (RQ) are as follows:

RQ1) To what extent are reader-perspective comments effective in bias detection?

RQ2) What is the impact of model size on bias detection performance?

RQ3) What kind of comment generation policies are most/more beneficial?

RQ4) What is best comment combination, and

RQ5) Whether a selector is necessary?

We conduct experiments on the news bias dataset BASIL (Fan et al., 2019) and the gender bias dataset BeyondGender (Luo et al., 2025) with Llamas and GPT4. Experimental results demonstrate the effectiveness and robustness of our method compared to previous methods (RQ1). Furthermore, we analyze the impact of different model sizes and comment-generation policies on the performance of bias detection (RQ2 & RQ3), providing insights into how LLMs can be utilized as readers in biased content analysis. In addition, the

optimal number of comment one and the Selector play a critical role (RQ4 & RQ5). Framing LLMs as active participants, our research offers a scalable and interpretable solution for identifying and mitigating bias in media content.

The main contributions are as follows:

1) A novel framework utilizing LLMs as Reader to generate comments for bias detection,

2) Effective and robust method on news bias and gender bias detection, and

3) Findings that emotion-related comments are generally more beneficial than value-related ones and that comments can be influenced by the reader's gender .

2 Related Work

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.

3 Framework

The framework consists of three roles: Reader, Selector, and Detector, as illustrated in Figure 1. Upon receiving data, the Reader generates comments based on the content. These comments are then evaluated by the Selector, which determines their usefulness. Finally, the Detector uses the original data along with the selected helpful comments to assess whether the data contains bias.

3.1 Reader-Perspective Design

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

General Perspective. This dimension examines the external and rational aspects of content (Stratton, 2021), 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. This dimension explores the internal and emotional responses elicited by content (Han and Arpan, 2017), focusing on:

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.

3.2 Component Design

Reader: Comment Generation. We employ Llama3.1-70B (Grattafiori et al., 2024) to produce reader-perspective comments from both general and individual perspectives. For each data, we instruct the LLM with "If yes, please specify" under the policies, as shown in Appendix Table 4 and 5. Specifically, we differentiate the reader's gender when commenting on gender bias data.

Selector: Comment Selection. We utilize generated comments to train a comment selector (BERT (Devlin et al., 2019)) capable of distinguishing between positive and negative comments. Initially, we record the LLM's prediction for each data. Then, we add a comment to the input and observe LLM's prediction on the comment-augmented data. A comment is labeled as positive if it changes an incorrect prediction to correct, and negative if it alters a correct prediction to incorrect.

Detector: Bias Detection. Provided with the orig-

LLM	BASIL Inf/ Lex / non		BeyondGender Sexism		Gender		Misogyny		Misandry	
	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC
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**	**0.85**	**0.75**	**0.40**	**0.38**	**0.85**	**0.76**	**0.18**	0.23
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
Existing SOTA	-	0.81	0.79	0.67	0.40	0.30	0.69	0.59	0.19	0.30
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

Table 1: Main results of baselines and comment-augmented models. The values are F1-scores and accuracy. The best results among open-source models are bolded. For Misandry, F1 is more reliable due to skewed label distribution.

inal data and selected positive comments, the Detector (an LLM) is instructed to detect bias in a zero-shot setting. The prompt is, for example, " $news$: + original_data + $comment$: + generated_comments + $Is\ the\ news\ biased?$ ". The word "news" is replaced with an appropriate term based on the data type.[1]

4 Experiment Settings

4.1 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". [2]

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 the original datasets are listed in Table 2. When training the Selector, we split 30% of the train set as the dev set.

4.2 Models

We evaluate the detection performance of Llama3.1-8B, Llama3.1-70B (Grattafiori et al., 2024), and GPT4 (OpenAI, 2023).

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).

Dataset	Label	Train	Test
BASIL	Inf	349	123
	Lex	138	32
	Non-bias	2,067	641
BeyondGender	Sexism	4381	485
	Gender	5,233	367
	Misogyny	5,233	367
	Misandry	5,233	367

Table 2: Statistic of the original datasets.

5 Results

5.1 Main Results

The main results, in Table 1, address **RQ1** (effectiveness) and **RQ2** (model size v.s. performance).

The effectiveness of our method is evidenced by substantial and consistent improvements in both F1-score and accuracy achieved by Llama-8B, reflecting significant gains in true positive and true negative rates. While accuracy on the BASIL dataset is similar to the existing SOTA, our approach offers greater explainability.

Regarding model size, larger LLMs like Llama-70B do not outperform smaller ones such as Llama-8B. However, with comment augmentation, both achieve comparable results (Llama-70B-AUG vs. Llama-8B-AUG), underscoring our method's effectiveness. 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, our method enables Llama-8B to perform on par with GPT-4 on the Sexism label and outperform it on the Misogyny label.

[1] Preliminary experiment shows that a simple prompt is better than a detailed one.

[2] 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
Llama-8B	0.73	0.61	0.32	0.33	0.72	0.62	0.16	0.43
Top-1	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
Existing SOTA	0.79	0.67	0.40	0.30	0.69	0.59	0.19	0.30

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.

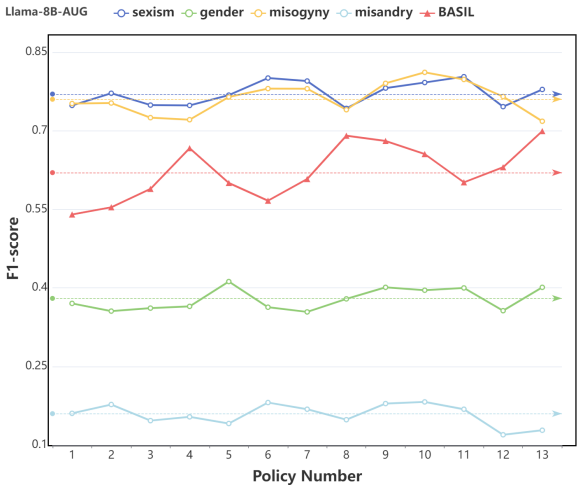

Figure 2: The F1-scores of each policy. The red line with triangles is BASIL; the purple, green, yellow, and blue lines with circles are Sexism, Gender, Misogyny, and Misandry, respectively. The dashed lines indicate the averages. Policy No.1-6 are general perspectives and No.7-13 are individual perspectives.

5.2 Policy Analysis

To answer **RQ3**, we compare each policy without comment selection, as shown in Figure 2.

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 best performance with policy No.11, 5, 10, 10, respectively. Moreover, Sexism, Misogyny, and Miandry have a similar trend, with policy No.6-7 and 9-11 above the average. Specifically, the gender difference between policy No.7 and No.8 leads to the performance gap (verse in Gender label), revealing the disparity of comments regarding reader's gender.

5.3 Comment Combination & Selector

To address **RQ4**, we conducted experiments with various comment combinations, as detailed in Table 3. Compared to the baseline Llama-8B, 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.

While only the Top-1 policy outperforms the existing SOTA across all labels, the selector enhances performance to a comparable level regardless of the comment combinations. This provides a confirmed answer to **RQ5** (necessity of Selector) and demonstrates the robustness of our method as well. Meanwhile, it suggests that the potential bottleneck of our approach may lie in the quality of the comments, which warrants further investigation.

6 Conclusion

In this work, we explore the potential of leveraging LLMs as readers to generate reader-perspective comments for bias detection. Through the design of general and individual comment generation policies, coupled with comment selection, our method demonstrates significant effectiveness and robustness in detecting bias across both news and gender bias datasets. Furthermore, the analysis of different model sizes and comment generation policies provide valuable insights into optimizing LLMs as readers in biased content analysis.

Limitations

The experimental results suggest that a key bottleneck may lie in the quality of the generated comments, as model 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.

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

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.

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 Thirty-Ninth AAAI Conference on Artificial Intelligence (AAAI 2025)*.

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.

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.

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 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.

B Computational Source

- **Comment generation**:
  Llama3.1-70B, 4 * A100.
  48 hours for BASIL and 72 hours for BeyondGender.

- **Selector training**:
  BERT, 1 * A6000.
  1 hour for BASIL and 4 hours for BeyondGender.

- **Bias Detection**:
  1) Llama3.1-8B, 2 * A6000.
  4 hours for BASIL and 4 hours for BeyondGender.
  2) Llama3.1-70B, 2 * A6000.
  4 hours for BASIL and 6 hours for BeyondGender.

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:	
7/8	How do you feel after reading the post if you are a man / woman? (e.g. angry, disappointed, excited, etc.)
9	Does the post have a positive, negative, or neutral tone?
10	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.