Pitfalls of Conversational LLMs on News Debiasing

Ipek Baris Schlicht[1,2], **Defne Altiok**[1], **Maryanne Taouk**[3], **Lucie Flek**[4,5,6]
[1]Deutsche Welle, Bonn/Berlin, Germany
[2]Universitat Politecnica de Valencia, Spain
[3]ABC News, Australia
[4]Conversational AI and Social Analytics (CAISA) Lab
[5]Bonn-Aachen International Center for Information Technology (b-it), University of Bonn
[6]The Lamarr Institute for Machine Learning and Artificial Intelligence

Abstract

This paper addresses debiasing in news editing and evaluates the effectiveness of conversational Large Language Models in this task. We designed an evaluation checklist tailored to news editors' perspectives, obtained generated texts from three popular conversational models using a subset of a publicly available dataset in media bias, and evaluated the texts according to the designed checklist. Furthermore, we examined the models as evaluator for checking the quality of debiased model outputs. Our findings indicate that none of the LLMs are perfect in debiasing. Notably, some models, including ChatGPT, introduced unnecessary changes that may impact the author's style and create misinformation. Lastly, we show that the models do not perform as proficiently as domain experts in evaluating the quality of debiased outputs.

Keywords: News Bias Correction, LLMs, Human Evaluation, Automatic Evaluation

1. Introduction

Biased news articles have the potential to significantly shape public opinion and discourse on various issues. Thus, professional news editors identify bias text spans in news articles before they are published. This task is particularly challenging, especially when editorial teams face constraints such as time-pressure and a lack of human resources.

Large Language Models (LLMs) have demonstrated outstanding performance even in the absence of labeled data, through zero-shot prompting. In many tasks, LLMs have surpassed the performance of the supervised models and have even employed as writing assistance (Shi et al., 2022; Zhang et al., 2023). In addition, conversational LLMs such as ChatGPT (OpenAI, 2022) and GPT4 (OpenAI, 2023) are user-friendly, making them accessible to non-technical experts like journalists who can use them without coding knowledge to aid in their tasks. As a result, many media companies have already begun experimenting ChatGPT for various journalistic tasks (Beckett, 2023). Limited studies have explored debiasing through text generation with conversational LLMs for the tasks such as hate speech (Plaza-del arco et al., 2023) and toxicity detection(Morabito et al., 2023). These studies explored zero-shot prompting with conversational LLMs. To the best of our knowledge, conversational LLMs have not been explored for news debiasing.

Standard evaluation metrics (Min et al., 2023) such as ROUGE require a reference text for measuring generated text quality and lacks explanatory evaluation. Morabito et al. (2023) established an evaluation protocol for automatically assessing LLMs' consistency in debiasing for toxicity detection by using Perspective API(per) as the evaluator. However, this protocol is limited to bias reduction and may not be suitable for the news domain. In the context of news bias, bias encompasses both overt bias such derogatory terms within text and latent biases that shape the language and framing of news stories (Recasens et al., 2013). As shown in Figure 1, news texts deemed biased may not contain toxicity but wording/phrasing could introduce bias. Hence, tools such as Perspective API could fail to quantify bias reduction. Furthermore, the debiased text might produce misinformation by changing context and factuality and altering the author's writing style. Therefore, there is a need for

President Donald Trump gave states and local governments the right to reject refugees, but instead of saying no, most state and local officials have blindsided the administration by opting in, according to two former officials familiar with the matter.
President Donald Trump allowed states and local governments the option to refuse refugees. However, according to two former officials familiar with the matter, most state and local officials have chosen to accept refugees.

Figure 1: Biased text where the usage "blind-sided" introduces bias by conveying a strong negative opinion about the actions of state and local officials and its GPT4 debiased version which doesn't contain toxicity according to Perspective API. Debiasing changed the facts and the context (factually incorrect statement highlighted in red, original version in blue).

The First Workshop on Language-driven Deliberation Technology (DELITE2024), pages 33–38

ID	Concept	Question
C1	**Correcting Bias**	**Does the model produce unbiased text? Grade 1-3** *The text is unbiased. (3)* *The text is partially biased. (2)* *The text is highly biased. (1)*
C2	**Preserving Information**	**Does the model change textual facts? Grade 1-3** *The text facts are still present. (3)* *Some facts are missing. (2)* *Facts are completely missing. (1)*
C3	**Preserving Context**	**Does the model change the meaning of text? Grade 1-3** *The meaning of the text is preserved.(3)* *The meaning of the text is partially preserved. (2)* *The meaning of the text is completely changed.(1)*
C4	**Preserving Language Fluency**	**Does the model produce grammatically correct text? Grade 1-3** *The text is grammatically correct. (3)* *The text has few grammar issues. (2)* *The text has many grammar issues. (1)*
C5	**Preserving Author's Style**	**Does the model harm the author's creativity? Grade 1-3** *No, the model did all necessary changes without harming author creativity. (3)* *The model corrected some of the texts that might hurt the creativity. (2)* *The model did unnecessary changes, and changed the text style. (1)*

Table 1: News editorial criteria for checking quality of debiasing.

evaluation criteria discerned editorial perspectives.

To address these issues, we investigate the following research questions (RQs): (1) How well do conversational LLMs perform debiasing in the context of the news domain according to editorial criteria? (2) Can conversational LLMs also serve as an evaluation tool for assessing the editorial quality of debiased articles?

Given the need for a domain-specific evaluation to assess the quality of conversational LLMs in news debiasing, we propose a set of evaluation criteria tailored to news editors. Since there is no publicly available news dataset for debiasing, we obtained text generations on a subset of the publicly available bias classification dataset using three popular conversational LLMs and a fine-tuned T5 (Raffel et al., 2020). Expert news editors from international media organizations ranked the models' outputs based on the editorial criteria. Additionally, we compared model outputs with expert assessments when the models were used as evaluation tools to check the quality of debiasing. Our results showed that despite conversational LLMs' proficiency in bias reduction, they sometimes generate misinformation and alter writing styles. Moreover, they can not assess debiased outputs as the experts do [1].

2. Related Works

The studies on media bias have primarily focused on two aspects: identifying biased text spans (Spinde et al., 2021; Hamborg, 2020; Lei et al., 2022) and detecting political bias in news articles (Chen et al., 2020) or media outlets (Baly et al., 2020). Only a few studies proposed methods for mitigating bias through article generation using transformer models. Among these studies, the earliest work by Pryzant et al. (2020) used BERT to identify subjective content and update the hidden layers of the model to generate unbiased text from Wikipedia. Lee et al. (2022) applied a summarization method on articles from various political leanings to neutralize news.

Plaza-del arco et al. (2023) and Morabito et al. (2023) explored the potential of zero-shot prompting with LLMs, respectively for hate speech detection and reducing toxicity in user comments. Additionally, Morabito et al. (2023) established an evaluation protocol for evaluating consistency of LLMs on debiasing in the context of toxicity detection. The authors used Perspective API as the evaluator tool which provides toxicity scores for comment moderation. However, the protocol is limited to only to bias reduction. Furthermore, is not applicable within the news domain as news articles may not exhibit a toxic tone, yet they can still contain biases favoring certain groups, which need to be addressed before publication. In our work, we design evaluation criteria taking into account journalistic perspective to measure quality of debiased sentences.

Recently, researchers have explored LLMs as evaluators for assessing the quality of text generation in various applications (Gao et al., 2024; Min et al., 2023) as an alternative solution to costly expert assessments. Motivated by this, we evaluate

[1]The code and the data are at https://bit.ly/3vGphbw

the conversational LLMs models as evaluators for assessing the quality of debiased sentences based on the journalistic criteria and compare them with our expert evaluation.

3. Methodology

We investigated three conversational LLMs for debiasing news sentences and paragraphs. Given sentences or paragraphs containing bias types such as epistemological, framing and demographic bias (Pryzant et al., 2020; Spinde et al., 2021; Recasens et al., 2013), the goal of the task was to generate an unbiased version of those sentences. The outcome of the sentences should be unbiased but other criteria should also be considered as important for news editors, such as preserving factuality, news' message, and not harming the authors' creativity, along with grammar changes.

3.1. News Editorial Criteria

As prior evaluation metrics are limited to news debiasing, we propose news editorial criteria. The editorial criteria were created during the implementation of BiasBlocker, which is a prototype AI-based news editor. [2].

The BiasBlocker team comprises experienced news editors and technologists from Deutsche Welle, ABC News and ARIJ. Since bias is a broad concept, to establish a common ground on the bias definitions and the corrections, the editors in the team created a codebook on bias types [3] and guidelines for debiasing based on the prior studies (Pryzant et al., 2020; Spinde et al., 2021; Recasens et al., 2013) and UN Guidelines [4]. Hence, the bias types we focus on are primarily framing, epistemological, and demographic bias.

We applied a pilot study on bias correction by using ChatGPT with the editors [5]. The editors spotted the issues and refined the expectations for the news editor. As outlined in Table 1, we distilled these expectations into five criteria for assessing the quality of models in the context of debiasing for news editing.

Essentially, the editors expected the model to effectively remove any text spans that introduce bias into the content. However, they also had the expectation that this must refrain from adding new facts or removing vital information, as this could produce misinformation. Furthermore, the model must ensure that the meaning of the text remains intact. The debiased text must also be grammatically correct. Lastly, especially for those articles of opinion pieces or analyses, the model must respect and preserve the author's writing style and creativity. Otherwise, the model could discourage less experienced authors and harm the communication of the news message.

Evaluation Dataset. Wiki Neutrality Corpus (WNC) (Pryzant et al., 2020) is the only publicly available dataset that contains biased samples and their debiased version by Wikipedia editors. Given that our research objective was to assess the LLMs in correcting bias within texts authored by news authors, WNC samples were not suitable for the evaluations. Therefore, we preferred the BABE dataset (Spinde et al., 2021) as the test set. BABE consists of sentences from news articles published by US publishers with different political leanings. Experienced media experts annotated the dataset; the dataset samples were labeled as biased or unbiased. The authors of the dataset provided two subsets. We chose the one annotated with more experts and randomly selected 50 biased sentences from this subset for the evaluations.

3.2. Debiasing Models

Baseline. As the baseline, we used the large version of T5 (Raffel et al., 2020). T5 is an encoder-decoder transformer that is pre-trained on a cleaned Common Crawl collection, incorporating a mixture of supervised tasks through multi-task learning. To adopt T5-large for the debiasing task, we used WNC as the training dataset. Given our constraints with computational resources, we fine-tuned the model using LoRA adaptation (Hu et al., 2022).

Conversational LLMs. We evaluated the popular conversational LLMs: ChatGPT (OpenAI, 2022) and GPT4 (OpenAI, 2023) from OpenAI, and Llama2-70b-chat (Touvron et al., 2023) which is an open source, popular alternative to ChatGPT. The models were adopted for conversational tasks using reinforcement learning with human feedback. In this way, individuals without technical expertise could easily interact with the models, making them suitable for integration into news organizations.

ChatGPT and Llama 2 are Autoregressive Language Models trained on large corpora from multiple sources from the web, with the objective of predicting the next word based on the preceding context. GPT4 is the advanced version of ChatGPT, capable of handling multi-modal input. While our task focused on textual input, we included GPT4 in our evaluations, because human evaluators preferred GPT4 outputs from various tasks over ChatGPT (OpenAI, 2023). We used prompts, which are shown in Table 2, for each of the conversational.

[2] https://bit.ly/4aJttWD
[3] https://bit.ly/49qcnvZ
[4] https://bit.ly/3PRks67
[5] https://bit.ly/43OMCnQ

Debiasing Prompt

Transform the following biased sentence into an unbiased sentence from a news article by removing any subjective language or discriminatory undertones without changing its semantic meaning:

Biased Sentence:

{{sentence}}

Unbiased Sentence:

Evaluator Prompt (shortened)

The input sentence from a news article is biased, it uses subjective language or discriminatory undertones. The other sentence was debiased by a language model. Your task is to compare two sentences based on the following journalistic criteria. For each question in the checklist, select your response from {1, 2, 3}.

The checklist is as follows:
1- Does the model produce unbiased text?
- If the text is unbiased, return 3
...

Do not explain your decisions.

Biased Sentence:

{{sentence}}

Model Output:

{{model_output}}

Checklist Answers:

Figure 2: Prompts for debiasing and evaluation. The full version of the evaluator prompt can be found at our source code.

4. Results

Although BABE contains the biased text spans along with the labels, the dataset does not have the corrected versions of the biased texts. Therefore, we could not directly apply the evaluation criteria to the samples. For this reason, two expert news editors from the team, as described in § 3.1, conducted the human evaluations voluntarily. Due to resource constraints, we split the models' results into two parts for both evaluators. Each part contains the results from each model. One editor ranked the samples which they were responsible for, by using a 3-likert scale. During the ranking evaluation, the editor marked the samples they were unsure about, made notes and applied fact-checking to address the C2 and C3. The other editor reviewed the ranked samples while checking the notes, marked samples and the fact-checked ones. The editors regularly engaged in discussions to reach a consensus on disagreements and uncertain cases. In total, we obtained 200 evaluations from the experts. Table 2 presents the frequency of ratings per criterion.

ID	Grade	T5	Llama2	ChatGPT	GPT4
C1	1	0.26	0.08	0.02	0
	2	0.40	0.06	0.14	0.38
	3	0.34	**0.86**	0.84	0.62
C2	1	0.1	0.4	0.26	0.2
	2	0.12	0.44	0.36	0.56
	3	**0.78**	0.16	0.38	0.24
C3	1	0.12	0.34	0.20	0.06
	2	0.06	0.4	0.48	0.68
	3	**0.82**	0.26	0.32	0.26
C4	1	0.34	0.1	0.02	0
	2	0.2	0	0	0.12
	3	0.46	0.9	**0.98**	0.88
C5	1	0.14	0.44	0.42	0.42
	2	0.08	0.46	0.56	0.5
	3	**0.78**	0.1	0.02	0.08

Table 2: The conversational LLMs are significantly better than the baseline at correcting bias and providing grammatically correct outputs (Student's T-test, p-value at 0.05), they have issues on preserving information, context and author's style.

RQ1: Debiasing Performance of the Conversational Models. The conversational LLMs proved better than the baseline for debiasing. Surprisingly, Llama 2 demonstrated comparative results even though ChatGPT has been known to outperform others in various tasks (Touvron et al., 2023). The researchers of Llama 2's training regime - that the factual sources were prioritized in training samples - might account for its competitive performance in this task. The conversational LLMs also exhibited more grammatical correctness than the baseline. Nevertheless, some LLMs changed phrases they considered biased, while others removed words or sentiments that could be considered confrontational or impolite, but are not actually biased towards any particular group. For instance, GPT4 changed 'When carrying a firearm, you have the ultimate power of force in your control' to 'When carrying a firearm, you have a significant level of potential force at your disposal'.

The conversational LLMs performed worse than the baseline model in preserving information and context. These models introduced unnecessary amendments to the generated texts. In some cases, even created hallucinations. This issue is not unique to this study and has been reported in related studies, especially in the case of ChatGPT being used for various tasks (Bang et al., 2023).Additionally, the news editors observed that Llama 2

introduced additional information not present in the input text, albeit factually accurate. For example, in a text mentioning 'Wilkens', the model replaced 'Wilkens' with 'Judge Wilkens'. The model may have memorized such information from its training dataset. This behavior by conversational models might harm the author's style.

ID	Llama2	ChatGPT	GPT4
C1	0.0666	-0.0489	0.1109
C2	-0.0145	0.0285	0.0018
C3	0.1971	0.0280	0.0263
C4	0.0597	0	0
C5	-0.0022	0.0454	-0.0413

Table 3: The disagreement between the conversational tools as an evaluator and the expert evaluation is high, according to Cohen's Kappa.

RQ2: Conversational LLMs as Evaluator: We obtained rankings from the conversational LLMs and compared them with the expert rankings. As shown in Table 3, there is a high disagreement between the models and the expert evaluations. Additionally, we observed that the models rated the criteria, such as preserving factuality, grammar, with the highest score. In contrast, the ratings by the experts for these criteria were low.

5. Conclusion

Through the editorial criteria, we showed that none of the conversational LLMs are perfect, even though they are good at debiasing. Specifically, they failed to preserve vital information and context, often leading to hallucinations. Employing these tools in a fully automatic editor can be dangerous, as they can create misinformation.

Memorization also surfaces as an important aspect of LLM behavior. For this reason, to ensure a fair evaluation of debiasing tasks across news articles from different periods, Media bias researchers need to create benchmark datasets containing samples from time periods that is not covered within the training data of LLMs.

The assessments by the models are not close to those by the experts. We plan to increase the size of our annotations and the number of annotators to build a benchmark dataset for a fine-grained analysis of the models' issues. We then investigate advanced methods for automating the evaluation criteria and incorporating them to adapt the models.

Ethical Considerations and Limitations

In this study, we assessed the efficiency of conversational LLMs in debiasing news articles, focusing solely on English samples from US Media. As a result, the generalizability of our conclusions to other languages and to media in other countries may be limited.

The dataset employed in this research paper is derived from publicly accessible sources and is peer-reviewed. During the evaluation process, we refrained from disclosing the identities of the article publishers to the participating news editors, thereby preventing any potential influence on their evaluations.

Acknowledgement

This research was partially funded by the JournalismAI Fellowship Programme 2024 of PolisLSE and vera.ai, which is co-financed by the European Union, Horizon Europe programme, Grant Agreement No 101070093. We also thank Kevin Nguyen, Saja Mortada, Khalid Waleed for their support.

References

Perspective api. https://perspectiveapi.com/. Accessed: 2024-02-15.

Ramy Baly, Georgi Karadzhov, Jisun An, Haewoon Kwak, Yoan Dinkov, Ahmed Ali, James Glass, and Preslav Nakov. 2020. What was written vs. who read it: News media profiling using text analysis and social media context. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3364–3374.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *CoRR*, abs/2302.04023.

Charlie Beckett. 2023. How newsrooms around the world use ai: a journalismai 2023 global survey. https://blogs.lse.ac.uk/polis/2023/06/26/how-newsrooms-around-the-world-use-ai-a-journalismai-2023-global-survey/.

Wei-Fan Chen, Khalid Al Khatib, Henning Wachsmuth, and Benno Stein. 2020. Analyzing political bias and unfairness in news articles at different levels of granularity. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 149–154, Online. Association for Computational Linguistics.

Mingqi Gao, Xinyu Hu, Jie Ruan, Xiao Pu, and Xiaojun Wan. 2024. Llm-based NLG evaluation: Current status and challenges. *CoRR*, abs/2402.01383.

Felix Hamborg. 2020. Media bias, the social sciences, and nlp: automating frame analyses to identify bias by word choice and labeling. In *Proceedings of the 58th annual meeting of the association for computational linguistics: student research workshop*, pages 79–87.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *ICLR*. OpenReview.net.

Nayeon Lee, Yejin Bang, Tiezheng Yu, Andrea Madotto, and Pascale Fung. 2022. Neus: Neutral multi-news summarization for mitigating framing bias. In *NAACL-HLT*, pages 3131–3148. 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.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Robert Morabito, Jad Kabbara, and Ali Emami. 2023. Debiasing should be good and bad: Measuring the consistency of debiasing techniques in language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4581–4597, Toronto, Canada. Association for Computational Linguistics.

OpenAI. 2022. Openai: Introducing chatgpt. https://openai.com/blog/chatgpt.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Flor Miriam Plaza-del arco, Debora Nozza, and Dirk Hovy. 2023. Respectful or toxic? using zero-shot learning with language models to detect hate speech. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 60–68, Toronto, Canada. Association for Computational Linguistics.

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*, volume 34, pages 480–489.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic models for analyzing and detecting biased language. In *ACL (1)*, pages 1650–1659. The Association for Computer Linguistics.

Shuming Shi, Enbo Zhao, Duyu Tang, Yan Wang, Piji Li, Wei Bi, Haiyun Jiang, Guoping Huang, Leyang Cui, Xinting Huang, Cong Zhou, Yong Dai, and Dongyang Ma. 2022. Effidit: Your AI writing assistant. *CoRR*, abs/2208.01815.

Timo Spinde, Manuel Plank, Jan-David Krieger, Terry Ruas, Bela Gipp, and Akiko Aizawa. 2021. Neural media bias detection using distant supervision with BABE - bias annotations by experts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1166–1177, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Yue Zhang, Leyang Cui, Deng Cai, Xinting Huang, Tao Fang, and Wei Bi. 2023. Multi-task instruction tuning of llama for specific scenarios: A preliminary study on writing assistance. *CoRR*, abs/2305.13225.