How LLMs Influence Perceived Bias in Journalism

Asteria Kaeberlein and Malihe Alikanhi
Northeastern University
{kaeberlein.c, m.alikhani}@northeastern.edu

Abstract

As the use of generative AI tools in journalistic writing becomes more common, reporters have expressed growing concerns about how it may introduce bias to their works. This paper investigates how the integration of large language models (LLMs) into journalistic writing, both as editors and independent 'authors', can alter user perception of bias in media. We show novel insights into how human perception of media bias differs from automatic evaluations. Through human evaluations comparing original human-authored articles, AI-edited articles, and AI-generated articles, we show that while LLMs rarely introduce new bias and often trend towards neutrality, this supposedly 'safe' behavior can have harmful impacts. This is most observable in sensitive human rights contexts, where the AI's neutral and measured tone can reduce the representation of relevant voices and present misinformation in a more convincing manner. Furthermore, we demonstrate the existence of previously unidentified patterns that existing automated bias detection methods fail to accurately capture. We underscore the critical need for human-centered evaluation frameworks in AI-assisted journalism by introducing human evaluations and contrasting against a state-of-the-art automated bias detection system.

1 Introduction

The increasing reliance on generative AI tools in journalism (New, 2024) has sparked substantial debate concerning the biases these models introduce or reinforce in news narratives. While AI offers efficiency in content creation, editing, and revision, it also introduces risks including hallucinations (Ji et al., 2023), stereotype reinforcement (Kotek et al., 2023), and inadvertent exclusion of marginalized voices (Gillespie, 2024). These significant flaws substantially damage the integrity and reliability of publishing companies that seek to introduce these tools to their work.

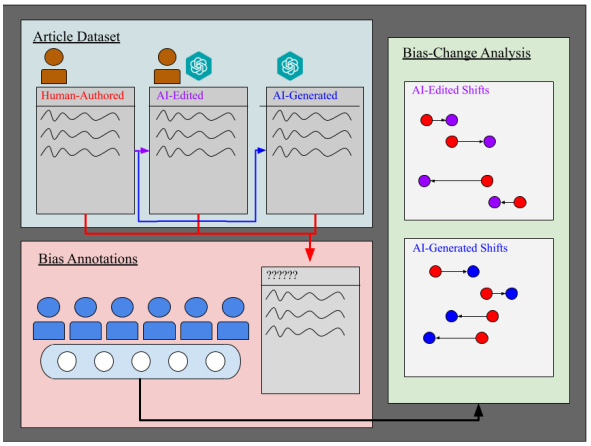

Figure 1: The process through which we evaluate how LLMs impact reader perception of articles. We first construct a dataset of human-authored articles and ask the AI to edit them or generate a new article on the same topic. These articles are then anonymized and passed to human annotators to rate bias. We then observe how the introduction of AI changes the perceived bias.

Existing studies primarily utilize automated metrics to quantify bias (Gallegos et al., 2024a), overlooking some subtle yet impactful biases that influence human readers. Such algorithmic approaches can inadequately address the complexity of bias in real-world contexts (Morini-Marrero et al., 2025). This is especially relevant in journalism where reader perception heavily affects the impact of published content (Sultan et al., 2024).

To address these limitations, our study examines how the inclusion of generative AI, as both editor and independent writer, affects human-perceived bias across political (Wei et al., 2023), racial (Lim and Pérez-Ortiz, 2024), and gender (Kotek et al., 2023) dimensions in journalism that AI often struggles with. We conduct human evaluations (as presented in Figure 1) of human-written, AI-edited,

and AI-generated news articles, complemented by comparisons against an advanced automated bias-detection model. Our analysis reveals critical insights: generative models consistently shift content toward neutrality (Askell et al., 2021), often inadvertently diluting critical perspectives on human rights and minority issues or leading to heightened perceived biases. Additionally, we demonstrate the inadequacy of existing automated tools in capturing these nuanced human perceptions, underscoring the necessity of human-in-the-loop evaluations.

This paper makes three primary contributions:

1. We contrast AI-generated and AI-edited articles to illustrate significant differences in how generative tools influence journalistic content (§ 4).

2. We expose gaps in current automated bias detection methods, emphasizing the need for more sophisticated human-aware evaluation techniques (§ 5).

3. We discuss how the practice of introducing LLMs to journalism can harm both consumers and producers of media (§ 6).

The remainder of the paper proceeds as follows: Section 2 reviews related work and describes our motivation for human analysis and defines the forms of bias studied. Section 3 presents the datasets and preprocessing techniques applied in this study. Section 4 describes the results and analysis of our human annotators, while Section 5 compares those results against a LLM-based evaluation.

2 Related Works

Human-Evaluation Numerous studies find that AI-based ratings align strongly with expert annotators in professional tasks such as literature analysis (Chiang and Lee, 2023) or nephrometry scores (Heller et al., 2022). However, performance drops when a model is tasked with more subjective or unprofessional evaluations (Morini-Marrero et al., 2025). For a task such as media evaluation, this is a major concern. The majority of consumers are non-experts and strict evaluations may fail to capture how readers actually consume articles.

The majority of people are not fair evaluators, particularly when it comes to media. Numerous features influence vulnerability to misinformation (Sultan et al., 2024). Furthermore, bias indicators can actually enhance a reader's own bias (Bruchmann et al., 2023). This suggests that evaluating how consumers perceive bias in media may not be a simple task. Due to this, many studies of media still rely on survey ratings, whether to confirm external ratings (French et al., 2025) or as an independent form of evaluation (Thurman et al., 2025).

As such, unlike most prior research into this topic (Gallegos et al., 2024a; Gillespie, 2024; Lim and Pérez-Ortiz, 2024), we elect to use human annotations instead of automatic evaluation. Media consumption is not a professional task and is heavily influenced by perception, suggesting that algorithmic evaluations may perform poorly. Ergo, unlike many prior studies that apply algorithmic evaluation to this task, we introduce annotators to capture more accurate interpretations.

Scope of Biases Studied We identify three primary forms of bias that are consistently demonstrated throughout prior research and are easily identifiable within papers by untrained annotators.

First, we discuss the political lean of an article. The traditional left-right dichotomy describes the amount of influence the government has over individual lives, where the left pushes for governmental interference to improve lives and the right argues for more freedom from government influence (Caprara and Vecchione, 2018). Unfortunately, due to the United States' extremely partisan politics, these factors have also been collated with a variety of shared beliefs over topics such as immigration (Rodrik, 2021) and abortion (Osborne et al., 2022), resulting in the traditional meaning of 'left' and 'right' no longer describing their corresponding political parties (Castles, 1982). Thus, we expect our primarily US-based participants will likely rate articles based on the positions of those parties. Generative AI has also demonstrated a slight left-leaning bias towards governmental interference as well as the democratic party (Motoki et al., 2024).

We also describe forms of racial bias. For the purpose of our survey, we consider forms of discrimination against underprivileged groups based on ethnicity or perceived ethnicity to be racial bias (Naicker and Nunan, 2023). While some human-written media demonstrates this form of bias explicitly (Langley, 2024), LLMs tend to avoid being so blatant (Haim et al., 2024). Instead, they match the overall trend towards more implicit bias, often through stereotyping or a lack of presentation.

Finally, we study gender bias. We consider gen-

der bias to be any form of discrimination based on gender. This category includes stereotypes of men and women, as well as discrimination of transgender individuals. Similar to racial bias, this has been very explicit in media (Moazami, 2023) while more neutral articles and generative models tend towards stereotyping (Kotek et al., 2023).

In contrast to prior research, we apply both an algorithmic evaluation method and layman annotator ratings. Since quantitative and qualitative evaluations of bias have proven effective in other areas (Gadiraju et al., 2023), we introduce similar elements to media analysis. We find that while automated systems are effective, they are unsuccessful at capturing some patterns in human evaluation.

3 Preprocessing and Data Collection

To identify articles that would show the forms of bias we discuss, we took a multi-step process of human annotation on the Webz.io News Dataset (Geva). This dataset contains 60k articles at time of access, collected from 2023 to 2024 across multiple countries. Articles are then grouped into 15 categories based on content. This includes topics such as “Environment", “Sport", and “Disaster". These are often grouped into ‘positive’ and ‘negative’ folders, referring to the position took on the matter.

Selecting Articles We first identify relevant subcategories that would cover human rights: “Politics", referring to news about politics, “Human Interest", which discusses aid to those in need, and “Social Issue", targeting specific wide-ranging moral concerns. 20 articles are randomly sampled from each of these categories, both positive and negative, as a part of our initial set. We add an additional article from each other category to demonstrate model performance on less biased writing.

Since we are asking a relatively small number of annotators to read and review these articles, we manually choose 11 that we believe give reasonable coverage of the various forms of bias measured. While this can limit generalizability, this enables us to collect multiple reviews of each article without requiring our annotators reading an unreasonable amount. We believe this tradeoff to be worthwhile as it reduces the impact of interrater reliability.

LLM-produced Articles and Preprocessing To generate our set of AI-edited articles, we direct the model to take the role of an editor at a publishing company and then provide it with the original text. This generally results in similar content to the original, though often with small changes. In order to create our set of AI-generated articles, we have the model take the role of a writer and prompt it to write its own paper on the same topic as the original. Prompts are run through GPT-4o, as it is a publicly available model being used in a variety of settings.

Preprocessing and Anonymization To ensure the articles would not be easily identified as AI-generated, we manually make slight corrections to characters and formatting choices the language model made. The majority of these are fixing the repeated use of asterisks and different characters being used as apostrophes. Indentation changes are also made to bring the articles more in line with those that were human-written. We do not make any modifications to change the generated textual content. Then, for each article, the raw text is placed within a document with a random name to prevent identifying AI-produced and human-produced papers.

3.1 Measuring Bias

To evaluate the bias, we create a survey form that asked participants to label each article on 5 metrics: gender bias, racial bias, political bias, article tone, and percieved factuality. We define bias as actions against protected attributes (Gallegos et al., 2024b). This means that there should not be harmful statements made against minorities, and that those groups should be represented in relevant discussions. What qualifies as a ‘harmful position’ may vary between different groups, but common examples are stereotypes and slurs.

We collect 6 participants through convenience sampling and provide them with the articles in a random order. For each article, they are asked to rate them on a bipolar Likert scale going from -3 to 3. Each metric also had a different description to provide the participants clarity on what the scale represented.

Likert scales are widely accepted in researching human behavior and have been used for psychology (Flückiger et al., 2016), education (Harpe, 2015), and bias analysis (Snipes et al., 1998). Additionally, it is also the structure used in the LLM-based method we compare human evaluations against (Watts et al., 2024). While that method elected to provide a 11-point Likert scale, we reduce it to a

7-point scale. Since these scales have issues with reliability and completion time (Dolnicar, 2021), we believe that reducing the number of options should make the process significantly easier for our annotators.

3.2 Defining Bias

When considering gender and racial bias, the scale is described as going from bias against a group to a neutral position to supporting the group in question. A major defining factor is the terms and representation of the relevant minorities. For example, discussing immigration without considering the racial minorities that are involved would be considered a '0' by default. However, referring to them with aggressive terms like 'invaders' tips that towards a '-3' rating. In contrast, a '+3' could be an article that uses quotes from the impacted population. Ultimately, this rating of bias considers whether relevant voices are heard and what tone is taken with those groups. It is worth observing that a rating of 0 is far closer to the metrics measured in prior studies. Such articles are representing the majority view and may rely on less harmful stereotypes.

Additionally, we measure political bias and article tone with a similar scale. Political bias is defined with '-3' to be left leaning and '+3' right-leaning. These values are selected to make the interface more intuitive, as '-3' is on the left. Article tone had '-3' represent a negative tone, such as frustration or disdain, while '+3' represents positive tones of celebration and excitement. This captures the attitude the article took on the content and the emotions that annotators felt it is trying to provoke.

Finally, we include a perceived factuality rating. This is meant to indicate whether citations were used and the content of an article is coherent and reasonably truthful. This metric demonstrates how likely the LLM is to introduce hallucinations to the articles. '-3' represents an unconvincing article lacking citations, '0' represents a seemingly realistic paper that lacks citations, or an unrealistic one that had citations. '3' represents a well-written paper that appears truthful and had relevant citations.

4 Results & Analysis

We observe in Table 1 that there are a few consistent patterns across almost every form of bias. Notably, we observe that the LLM makes substantial efforts to maintain a neutral position. This is reflected in originally negative articles moving more positive, while originally positive articles become more negative. Additionally, every metric had high variance. This largely stems from averaging both positive and negative together, creating a distance between the two and artificially inflating the variance. Additionally, Likert scales are rather unreliable (Dolnicar, 2021), so a significant degree of variance is to be expected. Annotators will naturally have a variance in rating depending on a number of factors, including their own background and how thoroughly they read the article. Despite these limitations, however, averaging still gives us a better understanding towards the general trend of the model's behavior.

Racial Bias Trends Towards Neutrality Let us now look to racial bias to assess those trends. We observe a positive average shift, which is a good thing. However, there's also a significant negative shift in articles that were already positive. This is an issue, since it suggests the removal of relevant voices from articles. While the effect is lessened when an AI is only editing, there is still a concerning shift away from representation. On the other hand, articles that were hostile to minorities are dialed down by a much larger degree.

Gender Bias Affected in Edits Similar patterns can be observed in gender bias. However, there is significantly less of a positive shift on average. Together with the insignificant shift of already-positive values, these suggest that the model will more often leave such voices untouched but will do less to correct bias against them. Also, to contrast with racial bias, gender bias shows a significant decrease between AI-edited articles and AI-generated ones. This highlights why analyzing how LLM's write collaboratively is important. The model shows significantly less bias against gender minorities when relying on another's work, but the bias observed in prior papers (Kotek et al., 2023) is observable when it is told to generate its own paper.

Political Bias is Left-leaning We are also able to recreate the slight left-leaning position LLM's take (Motoki et al., 2024) politically. Particularly in AI-generated text, it generally has a negative trend. This is especially noticeable in already left-leaning articles where the AI often made only small changes in position when generating its own article.

Form of Bias	AI Action	Average Shift	Variance	Negative Shift	Positive Shift
Racial	Edited	0.312	0.983	0.75 (0.40)	-0.125 (0.69)
Racial	Generated	0.23	0.695	1.0 (0.25)	-0.417 (0.42)
Gender	Edited	0.174	1.252	0.875 (0.38)	0.0 (0.70)
Gender	Generated	-0.024	1.236	0.75 (0.38)	-0.333 (0.37)
Political	Edited	-0.008	1.098	0.708 (0.02)	-0.75 (0.07)
Political	Generated	-0.106	0.799	0.25 (0.11)	-1.0 (0.14)
Tone	Edited	0.052	0.648	0.575 (0.19)	-1.167 (0.08)
Tone	Generated	0.488	0.562	0.583 (0.09)	-0.167 (0.51)
P. Factuality	Edited	0.297	0.714	1.25 (0.13)	0.3 (0.83)
P. Factuality	Generated	-0.405	0.736	0.75 (0.12)	-0.917 (0.003)

Table 1: A summarization of the change in ratings between human-written data and the ratings of those articles after AI intervention. The rating of an article is the average of its ratings from each reviewer. The average shift and variance are of the aggregation across all of the articles. The negative and positive median shift represent how significantly values that were originally negative or originally positive shifts after introducing AI. A visualization of these metrics can be observed in 2. These shifts often have the opposite sign of the original value as the AI moves towards a neutral position. These median shifts also include the p-value from the Wilcoxon Signed-Rank test to show how statistically reliable the shifts are.
In general, the model pushes most forms of bias towards a more neutral position. This can be observed in the negative median shift being positive, and vice versa for the positive median shift. Additionally, AI-edited papers often show significantly less bias than purely AI-generated papers. Finally, while AI-generated articles are perceived as less factual, AI-edited articles are actually seen as more factual than their originals.

Article Tone is Positive Tone of 'voice' also demonstrates an interesting pattern, albeit in the opposite direction according to our scale. AI-Edited articles show changes towards neutrality, but those the LLM generated are overwhelmingly more positive. Positive ratings - polite or enthusiastic - are changed minimally while hostile articles are made more neutral.

Perceived Factuality Increases in Edits Most interestingly, perceived factuality demonstrates a very interesting phenomena. As expected, articles that are purely AI-generated show a significant negative shift in factuality, seeming less truthful and less reasonable. However, those that are simply edited by the model show a positive shift even when they were originally rated positively. This highlights an interesting pattern where the model is able to make poorly phrased arguments sound more reasonable to our participants.

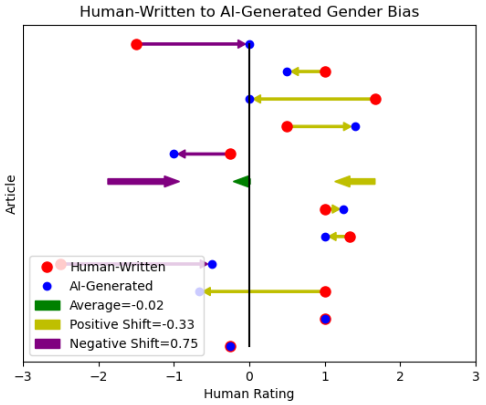

Figure 2: The Gender Bias ratings of AI generated articles corresponding to the fourth row of Table 1. The negative shift is measured by the median shift on the left, where values were originally negative. Similarly, the positive shift is measured by the median shift on the right.

4.1 Tone Impacts Perception

One of the concerns we raised regarding prior studies is that algorithmic analysis may fail to capture reader experience. Our results suggest that the way a statement is phrased can impact how users interpret it. This can strongly be observed in Figure 3. While no causation is inherently implied by the upwards trend, polite articles are often perceived as more truthful. Notably, this also held true to a lesser degree with AI-generated articles. Perceiving warmth and politeness as believable is an observed behavioral phenomena (Demeure et al., 2011), suggesting that the pattern we observe likely has an element of causality.

This has significant implications towards reader experience and evaluation of LLMs. There is strong correlations between demographic bias and

tone in both human-written and AI-processed articles. This provides us solid reason to believe that while LLMs may be able to decrease bias, they seem good at making potentially harmful information sound believable. This is a significant potential harm that has no existing evaluative measure. Additionally, this shows that current algorithms are not equipped to properly capture how humans interpret media.

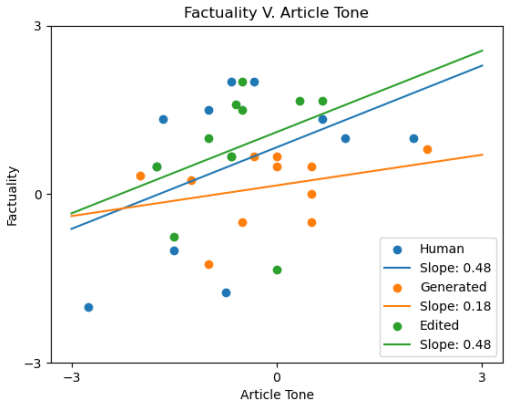

Figure 3: The factuality participants rated, plotted against the perceived tone of the article. Both human-written and AI-edited articles show a significant upwards trend.

5 Contrasting Automated Results

While most metrics to measure bias are well-defined and therefore limited, there do exists tools that may be capable of recognizing the nuances that humans do. LLMs have demonstrated incredible performance at complex tasks, so it seems reasonable to think they would do the same here. As such, we turn to one such example (Watts et al., 2024) that has been used as a tool to aid people in assessing media bias (Wang et al., 2025). We made slight modifications to the prompts provided by this architecture to bring them in line with our own scales, most notably shifting from an 11-point bipolar scale to a 7-point one. This architecture is then provided the same information as the human annotators. Additionally, we modify the 'political lean' prompt to measure gender and racial bias by replacing examples appropriately.

As observed in Table 2 the performance is disappointing. The LLM-based tool is unable to produce the same ratings our human annotators did. While it provides quick and easy assessments of bias, it does not appear to be usable if we wish to simulate human analysis. Unfortunately, this leaves us without an efficient way to assess the complex interplay of factors that go into media consumption.

In addition to quantitative failures, we also observe some qualitative patterns that provide some reasoning for the failure of the model. Interestingly, the LLM may take a similar approach to other algorithmic patterns and try to measure the occurrences of words. This can be observed in the way it rates an article describing 'sexism, racism, homophobia, and transphobia' as 'arbitrary' and fictional equal to one that is discussing how those behaviors are wrong. This is emphasized by its inability to understand article tone, rating the vast majority of articles significantly lower than the human annotators.

Furthermore, the model rates almost every neutral article to be 'left-leaning'. We speculate that this could be a demonstration of the model's own bias (Motoki et al., 2024) interpreting neutral and truthful articles as in-line with its own 'beliefs', and would heavily suggest further research into this pattern. Despite this shortcoming, it is fairly accurate in classifying right-leaning articles in both over-all political position and the degree of that position. This could obviously be harmful to right-leaning users of such a tool who would be disinclined to interact with neutral and truthful articles (Rhodes, 2022).

6 Discussion

Overall, we find that existing methods for estimating the bias AI introduces to articles are effective for their purposes, but insufficient for assessing how LLMs impact journalism. Multiple factors contribute to how humans view media that the over-simplified automated methods fail to capture. A major example of this is less direct forms of bias, which are often still present in neutral-tone articles. While the most common form of bias shown is tyranny of the majority, our work demonstrates that measuring only that form of bias does not capture the bigger picture. Additionally, we show that no complex architecture is currently able to simulate human annotations.

We also introduce the concept of evaluating both AI-generated and AI-edited articles. This is particularly important, as many areas where these techniques might be adopted are unlikely to fully forego human authors. The New York Times has a team working with AI, but they retain human editors to

	F1-score	Accuracy	Precision
Racial Bias	0.274	0.394	0.210
Gender Bias	0.240	0.303	0.205
Political Bias	0.315	0.212	0.728
Article Tone	0.105	0.182	0.077

Table 2: How accurately the LLM is able to classify bias. In general, it shows very poor performance across all areas, suggesting that the model is unable to accurately approximate human ratings.

review their publications instead of blindly trusting AI (New, 2024). Similarly, tools like Grammarly are adopting generative models to assist with editing, meaning that these will end up in the hands of students. We have highlighted how there is a substantial gap between assessing generative AI in a vacuum and a very simple collaborative setting where it makes small modifications to an article. Even with this reduction of a problem, we observe substantial differences in behavior. As such, it is highly important that future works consider how journalism is integrating LLMs when researching their impact.

We also highlight how LLMs are capable of convincing readers that their statements are more factual by being more polite than human writers. This could potentially be leveraged to spread misinformation, particularly in a collaborative setting. While we did not observe significant sycophantic behavior in our study, past studies have shown that repeated interactions better enable users to convince models of untrue or biased information (Sharma et al., 2023). This suggests it would be trivial to create an environment where a model is able to convincingly present misinformation in a collaborative setting. This provides the advantage of being more polite and thus seen as more truthful, while lacking actual factuality.

This has a number of applications that could cause significant harm. First, and most obvious, is it encourages the spread of misinformation that has drastically risen within the last few years. A significant upwards trend has been noted in politics (Shao et al., 2018), vaccinations (Nsoesie et al., 2020), and climate change (Farrell, 2019). Misinformation has also been leveraged against a number of demographics (Billard, 2023; Shimizu, 2020; Ndumu and Orie Chuku, 2023). The LLM trending toward neutrality is mirrored by existing behaviors in the real world (Georgi, 2025), which suggests that the introduction of LLMs to journalism could bolster existing discrimination through misinformation.

Second, it enables the framing of unequal perspectives as equal. This could be leveraged in a similar manner as some youtube 'debate' channels (Jubilee, 2019; Mohammed, 2019), which present unequal perspectives on equal footing. Such behaviors have led to the resurgence of conspiracy theories (Pannofino, 2024). While this parallels our discussion regarding misinformation, conspiracy theories may not be explicitly untrue. However, enabling conspiracy theories encourages complacency (Kendzior, 2022), and such 'free-thinking' perspectives are often directly linked with various forms of bigotry (Beringuy and Alvim, 2024).

Finally, it enforces the bystander effect. While a neutral position can present atrocities as normal, such as institutional slavery (Stevenson, 2022), strongly positioned media is able to sway people to assist each other (Javor, 2023). The neutral tone shown by the AI moves readers away from beneficial mutual assistance and towards simply standing aside, because that is what has been presented as 'normal'.

Overall, these potential harms indicate that LLM use in journalism could have significant negative impacts. The more neutral tones the model provides have been observed in other forms of media to cause significant harm to both minority demographics and consumers. This heavily suggests that we need greater work into how we apply AI to journalism. While simple methods are decent for simple tasks, they fail to capture the myriad of ways that bias could be introduced.

7 Conclusion

In this work, we highlight how generative AI can impact bias and perceived truth in journalism, and compare existing automated evaluations against human evaluations. We observe patterns that well-defined methods are unable to capture, such as the more neutral tone of LLMs being more persuasive to users, independent of bias. Finally, we show that LLMs currently lack the capacity to mirror

human evaluations and highlight how no existing method is properly able to evaluate how biased LLM-generated articles are. Furthermore, we elaborate how these significant faults can be harmful to various groups. Overall, we demonstrate that introducing LLMs to journalism can be dangerous, and show how the risks current models pose can be unintuitive and may be overlooked by manual review.

8 Limitations

This study had a few significant limitations. Our participants were of limited demographics, as we use convenience sampling, and due to the limited number of annotators we only test 11 human-written articles. While they address a fairly wide spread of topics, they are certainly not universal. A larger scale survey could find that some results we show do not hold on a grander scale. These participants were also very US-centric, which can heavily influence media interpretation and perception of biases. However, this further emphasizes how important human annotation can be, since bias-rating systems are unlikely to capture local interpretations of media.

Additionally, we only test GPT-4o for editing and generation. We believe that since GPT-4o showed less bias compared to other models in prior studies, it should do the same under human evaluation. However, it is worth considering that GPT has shown a tendency towards safety and neutrality. This contributes heavily towards the patterns we observe here and this may not be reflected in other models. However, this neutrality is intended to maintain factuality (OpenAI, 2025), something that should be a goal for tools designed to aid in journalistic tasks.

Finally, we did not test other predictive models to compare with our ratings. While we studied one of the more recent techniques that uses a state-of-the-art LLM (Watts et al., 2024) and has been tested by consumers (Wang et al., 2025), other simpler architectures have achieved very good performance at matching human rating scores for specific forms of bias. It is feasible that a combination of various architectures could achieve significantly better performance at matching human ratings.

Ethics Statement

A significant ethical concern that emerges from this paper is the limitations of convenience sampling. This could easily be interpreted as placing a heavy emphasis on a specific demographic to serve as 'judges'. However, this actually further emphasizes the need for future research to consider human interpretation. If our work does not generalize, that shows the variety of ways media can be interpreted meaning automated evaluations lack accurate interpretations.

Another significant ethical concern that emerged from this research was that many articles presented to annotators were extremely negative towards individuals of various demographics. Some covered sensitive topics such as sexual assault and mental health. Others included racism, transphobia, and homophobia. This makes recruitment very difficult and contributed to our smaller set of annotators. As such, expanding these results could be difficult. For future works to be ethical, they must properly present the potential topics annotators may need to deal with and ensure they are doing so in safe environments.

References

2024. Introducing the a.i. initiatives team. *The New York Times Company*.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.

Adriano Beringuy and Leandro Guimarães Marques Alvim. 2024. Hate, prejudice and conspiracy theories: The reality from the ideological perspective of brazilian imageboard users. *Journal of Language Aggression and Conflict*.

Thomas J Billard. 2023. "gender-critical" discourse as disinformation: Unpacking terf strategies of political communication. *Women's Studies in Communication*, 46(2):235–243.

Kathryn Bruchmann, Subramaniam Vincent, and Alexandra Folks. 2023. Political bias indicators and perceptions of news. *Frontiers in Psychology*, 14:1078966.

Gian Vittorio Caprara and Michele Vecchione. 2018. On the left and right ideological divide: Historical accounts and contemporary perspectives. *Political Psychology*, 39:49–83.

Francis Geoffrey Castles. 1982. Left-right political scales: Some'expert'judgements.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*.

Virginie Demeure, Radosław Niewiadomski, and Catherine Pelachaud. 2011. How is believability of a virtual agent related to warmth, competence, personification, and embodiment? *Presence*, 20(5):431–448.

Sara Dolnicar. 2021. 5/7-point "likert scales" aren't always the best option their validity is undermined by lack of reliability, response style bias, long completion times and limitations to permissible statistical procedures.

Justin Farrell. 2019. The growth of climate change misinformation in us philanthropy: evidence from natural language processing. *Environmental Research Letters*, 14(3):034013.

Christoph Flückiger, Hansjörg Znoj, and Andreea Vîslă. 2016. Detecting information processing bias toward psychopathology: Interpreting likert scales at intake assessment. *Psychotherapy*, 53(3):284.

Aaron M French, Veda C Storey, and Linda Wallace. 2025. The impact of cognitive biases on the believability of fake news. *European Journal of Information Systems*, 34(1):72–93.

Vinitha Gadiraju, Shaun Kane, Sunipa Dev, Alex Taylor, Ding Wang, Emily Denton, and Robin Brewer. 2023. " i wouldn't say offensive but...": Disability-centered perspectives on large language models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 205–216.

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

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024b. Bias and fairness in large language models: A survey. *Preprint*, arXiv:2309.00770.

F Richard Georgi. 2025. The violence we unsee: Human rights erasures in supply chain capitalism. *International Political Sociology*, 19(1):olaf002.

Ran Geva. Webz.io news dataset repository.

Tarleton Gillespie. 2024. Generative ai and the politics of visibility. *Big Data & Society*, 11(2):20539517241252131.

Amit Haim, Alejandro Salinas, and Julian Nyarko. 2024. What's in a name? auditing large language models for race and gender bias. *arXiv preprint arXiv:2402.14875*.

Spencer E Harpe. 2015. How to analyze likert and other rating scale data. *Currents in pharmacy teaching and learning*, 7(6):836–850.

Nicholas Heller, Resha Tejpaul, Fabian Isensee, Tarik Benidir, Martin Hofmann, P Blake, Zachary Rengal, Keenan Moore, Niranjan Sathianathen, Arveen Adith Kalapara, et al. 2022. Computer-generated renal nephrometry scores yield comparable predictive results to those of human-expert scores in predicting oncologic and perioperative outcomes. *The Journal of urology*, 207(5):1105–1115.

Andrew Javor. 2023. *Sympathetic Bystanders: The Dissemination of the Holocaust and Reactions by Gentile Britons, 1939-1945*. Ph.D. thesis, University of Guelph.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

Jubilee. 2019. Flat earthers vs scientists: Can we trust science? | middle ground.

Sarah Kendzior. 2022. *They knew: How a culture of conspiracy keeps America complacent*. Flatiron Books.

Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of the ACM collective intelligence conference*, pages 12–24.

Katherine Langley. 2024. Racism, reporting and the 'new plan for immigration', an analysis of uk media and legal and practical implications. *Journal of Immigration, Asylum & Nationality Law*, 38(1):50–66.

Serene Lim and María Pérez-Ortiz. 2024. The african woman is rhythmic and soulful: An investigation of implicit biases in llm open-ended text generation. *arXiv preprint arXiv:2407.01270*.

Sahar Moazami. 2023. Legalizing transphobia: How the anti-gender movement utilizes the law to uphold anti-trans hate. *Cal. W. Int'l LJ*, 54:179.

Shaheed N Mohammed. 2019. Conspiracy theories and flat-earth videos on youtube. *The Journal of Social Media in Society*, 8(2):84–102.

Sandra Morini-Marrero, Jose M Ramos-Henriquez, and Anil Bilgihan. 2025. Analyzing the concordance and consistency of ai and human ratings in hospitality reviews. *Journal of Hospitality and Tourism Technology*.

Fabio Motoki, Valdemar Pinho Neto, and Victor Rodrigues. 2024. More human than human: measuring chatgpt political bias. *Public Choice*, 198(1):3–23.

Ramona Naicker and D. Nunan. 2023. Racial bias.

Ana Ndumu and Nenna Orie Chuku. 2023. Anti-black racism, anti-immigrant sentiment, and misinformation: A recipe for profound societal harm. *Proceedings of the Association for Information Science and Technology*, 60(1):677–680.

Elaine Okanyene Nsoesie, Nina Cesare, Martin Müller, and Al Ozonoff. 2020. Covid-19 misinformation spread in eight countries: exponential growth modeling study. *Journal of medical Internet research*, 22(12):e24425.

OpenAI. 2025. Openai model spec.

Danny Osborne, Yanshu Huang, Nickola C Overall, Robbie M Sutton, Aino Petterson, Karen M Douglas, Paul G Davies, and Chris G Sibley. 2022. Abortion attitudes: An overview of demographic and ideological differences. *Political Psychology*, 43:29–76.

Nicola Luciano Pannofino. 2024. The "global" deception: Flat-earth conspiracy theory between science and religion. *Genealogy*, 8(2):32.

Samuel C Rhodes. 2022. Filter bubbles, echo chambers, and fake news: How social media conditions individuals to be less critical of political misinformation. *Political Communication*, 39(1):1–22.

Dani Rodrik. 2021. Why does globalization fuel populism? economics, culture, and the rise of right-wing populism. *Annual review of economics*, 13(1):133–170.

Chengcheng Shao, Pik-Mai Hui, Lei Wang, Xinwen Jiang, Alessandro Flammini, Filippo Menczer, and Giovanni Luca Ciampaglia. 2018. Anatomy of an online misinformation network. *PloS one*, 13(4):e0196087.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, et al. 2023. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.

Kazuki Shimizu. 2020. 2019-ncov, fake news, and racism. *The lancet*, 395(10225):685–686.

Robin L Snipes, Sharon L Oswald, and Steven B Caudill. 1998. Sex-role stereotyping, gender biases, and job selection: The use of ordinal logit in analyzing likert scale data. *Employee Responsibilities and Rights Journal*, 11:81–97.

Mark Stevenson. 2022. Hidden in plain sight: the bystander effect and the mobilisation of modern slavery whistleblowing. *Supply Chain Management: An International Journal*, 27(1):128–139.

Mubashir Sultan, Alan N Tump, Nina Ehmann, Philipp Lorenz-Spreen, Ralph Hertwig, Anton Gollwitzer, and Ralf HJM Kurvers. 2024. Susceptibility to online misinformation: A systematic meta-analysis of demographic and psychological factors. *Proceedings of the National Academy of Sciences*, 121(47):e2409329121.

Neil Thurman, Sally Stares, and Michael Koliska. 2025. Audience evaluations of news videos made with various levels of automation: A population-based survey experiment. *Journalism*, 26(1):3–23.

Jenny S Wang, Samar Haider, Amir Tohidi, Anushkaa Gupta, Yuxuan Zhang, Chris Callison-Burch, David Rothschild, and Duncan J Watts. 2025. Media bias detector: Designing and implementing a tool for real-time selection and framing bias analysis in news coverage. *arXiv preprint arXiv:2502.06009*.

Duncant Watts, Jeanne Ruane, David Rothschild, Anushkaa Gupta, Yuxuan Zhang, Jenny Wang, Amir Tohidi, Samar Haider, Calvin Isch, Timothy Dorr, and et al. 2024. Methodology: Media bias detector.

Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V Le. 2023. Simple synthetic data reduces sycophancy in large language models. *arXiv preprint arXiv:2308.03958*.