CITATION INEQUITY IN THE AGE OF AI: RISKS AND GAPS IN CURRENT RESEARCH

Mohammad Ausaaf Nasir
ausaafnasir24@gmail.com

ABSTRACT

Large language models like ChatGPT and Claude are changing how academic writing is done. They can help generate summaries, suggest citations, and even write whole sections of scientific text. But as these tools become more common, we need to ask how they might shape what gets cited, and whose work is left out. This paper looks at the risk that AI systems trained on existing publication patterns may reinforce the same citation biases that already exist in science, including those based on geography, prestige, and visibility. Drawing on recent studies, we show how citation is not just a technical feature of writing but a form of power that determines which voices are remembered. We also share a small experiment using GPT-4 to illustrate how these patterns appear in practice. Finally, we outline some concrete steps that researchers, developers, and institutions can take to promote fairer, more inclusive citation practices as AI becomes more involved in scholarly work.

***K*eywords** Citation Bias · Large Language Models · Epistemic Fairness · NLP fairness · AI-generated scientific writing

1 Introduction

Large language models like ChatGPT and Claude are being used more often in academic writing not just for grammar or summarization, but for drafting abstracts, generating literature reviews, and even suggesting citations. They're fast, fluent, and convenient, which explains the enthusiasm. But as they become more common, we should be asking: how might they shape what gets written, and who gets cited? This isn't just a technical concern. Citations are deeply tied to visibility and recognition in science. They affect careers, funding, and even which ideas survive. If AI tools start reinforcing the same dominant sources over and over like high-impact journals, elite universities, well-cited Western authors, then the diversity of perspectives in science could shrink over time. Some researchers are already worried about this. Peterson [1] talks about "knowledge collapse," where repeated AI use leads to convergence around average, high-probability ideas, erasing smaller or less mainstream contributions. Holtzman et al. [2] found that language models tend to produce high-probability text that sounds natural but carries little new information, a pattern that could be dangerous in science, where novelty and nuance are essential. Birhane et al. [3] argue that fairness in AI needs to include epistemic awareness, not just demographic parity.

Despite this, there's very little work focusing on citation fairness in AI-assisted scientific writing. Most fairness research in NLP is about downstream tasks like sentiment analysis or resume screening, not academic knowledge production. This paper offers a short survey of what we do know about citation bias, how LLMs might reinforce it, and what current AI research is missing. To understand how serious this risk is, we first need to examine how citation practices are already shaped by deep social and structural biases. These patterns did not begin with AI, but AI may accelerate and conceal them.

2 The Hidden Politics of Citation: How Academic Power Shapes Knowledge

Citations are often treated as neutral indicators of influence or quality. But in reality, citation practices are shaped by social structures, power, geography, language, prestige. These forces determine whose work gets seen, cited, and remembered. For example, Caplar et al. [4] analyzed citation patterns in astronomy and found that papers authored by women received significantly fewer citations than those authored by men, even after controlling for variables like

seniority and journal impact. Geographic bias is also widespread. Wuestman et al. [5] showed that scientific citations often cluster within the same regions or institutions, particularly disadvantaging scholars from the Global South, even when their work appears in top journals. Prestige bias plays a similar role: papers from elite universities or famous authors are more likely to be cited, sometimes regardless of content quality (Frachtenberg & McConville) [6]. These patterns aren't just academic curiosities, they have real consequences. Citation counts are used to evaluate researchers, decide promotions, and allocate funding. As a result, biased citation networks can perpetuate existing inequalities in science. As Pereira [7] argues, even efforts to "correct" citation imbalances can unintentionally reinforce hierarchies if they focus on surface level metrics rather than deeper structural change.

What this tells us is that citations are not just a way to acknowledge prior work. They're a form of epistemic power. They shape which ideas survive and which ones get buried. If AI tools are going to participate in this process, by generating citations, summarizing literature, or assisting in writing, then we need to understand how those tools might reproduce the same biases or make them worse. Multiple empirical studies have demonstrated how citation practices embed long standing academic inequalities. Table 1 summarizes some of the clearest findings, showing how bias emerges through gender, geography, institutional prestige, and feedback mechanisms in academic publishing.

Type of Bias	**Study**	**Scope / Field**	**Key Finding**
Gender bias	Caplar et al. [4]	Astronomy	Women-authored papers receive fewer citations than men's papers, even after controlling for seniority and journal impact.
Geographic bias	Wuestman et al. [5]	Multidisciplinary	Citations tend to cluster regionally, with scholars in the Global South receiving fewer citations even in top-tier journals.
Prestige bias	Frachtenberg & McConville [6]	Computer Systems	Peer review and citations favor submissions from elite institutions regardless of content quality.
Structural bias	Pereira [7]	Sociology	Even corrective citation efforts can reinforce academic hierarchies when surface-level metrics are applied uncritically.
Structural bias	Sinatra et al. [8]	Network Science	Citation impact results from both ability and cumulative advantage, creating long-term visibility gaps among scholars.

Table 1: Summary of Empirical Studies on Citation Bias in Scientific Publishing

These biases are reinforced by how academic influence is measured. Citation-based indicators like the h-index and journal impact factor are widely used to judge research quality, but they often reflect institutional privilege more than intrinsic merit. Scholars from the Global South or working in emerging areas are less likely to publish in high-impact venues, and as a result, receive fewer citations. When these metrics become part of LLM training data, they don't just reflect scientific patterns, they encode structural inequalities. Bender et al. [9] caution against training language models on unfiltered corpora, noting that large models may reproduce and even amplify dominant ideologies embedded in the data. They describe this as the danger of "stochastic parrots", systems that fluently generate familiar patterns without understanding or scrutiny. This concern overlaps with earlier work by Winner [10], who argued that technologies are never entirely neutral. The tools we design, including citation algorithms and writing models, inevitably reflect human values and hierarchies. When academic AI systems are trained on data shaped by prestige and volume, they risk perpetuating a narrow vision of what counts as legitimate knowledge.

This cycle is not just theoretical. It can be traced step by step. Language models are trained on scientific corpora that reflect historical citation trends. These corpora are shaped by metrics that reward already dominant authors and institutions. Once trained, the model produces outputs, literature reviews, summaries, suggested references, that tend to reflect the same dominant sources. When these outputs are used in academic work, they contribute to new texts that cite the same kinds of sources again. Eventually, these new AI-assisted papers become part of the training data for future models, creating a loop where inequality is quietly reinforced.

If citation is already biased, then what happens when machines begin generating it? This is the question that motivates our next section, where we turn to the role of large language models in shaping citation behavior.

3 When Machines Write Science: The Role of LLMs in Citation Patterns

As large language models (LLMs) enter academic workflows, they're not just helping with phrasing or summarizing, they're increasingly influencing what gets cited. Tools like ChatGPT or Claude can now suggest references, generate literature reviews, and even insert citations directly into drafts. But unlike human authors, these models don't make choices based on deliberate reading or intellectual judgment. They rely on statistical patterns learned from massive text corpora. That raises serious questions about whose voices these systems are most likely to amplify. Huang et al. [11] show that hallucinated citations generated by LLMs often reflect existing citation distributions, favoring well-known authors, high-impact journals, and mainstream topics. In other words, when language models guess, they guess in line with dominant trends. This means marginalized scholars or niche research areas are less likely to appear, even in fabricated outputs. Meanwhile, Liang et al. [12] found a sharp increase in AI-generated content in recent publications and noted that much of it reused similar framing and citation styles, suggesting a narrowing of variation over time. These patterns mirror concerns raised by Peterson [1] and Holtzman et al. [2], who warn that repeated use of generative AI may push scientific writing toward high-probability, low novelty output. The result could be what Peterson calls "knowledge collapse", a convergence toward predictable, averaged citations and concepts, with less room for theoretical risk or minority perspectives.

What's especially troubling is that these biases are often invisible to users. An author using ChatGPT to help write a paragraph may not realize that the suggested citations are drawn from a narrow, skewed sample. There is no warning label that says: "This list favors elite institutions" or "This misses Global South contributions." As a result, the use of AI might reinforce precisely the citation gaps that academia has struggled to fix for decades.

To see how this plays out in practice, we conducted a small test. GPT-4 was asked to generate short literature reviews on different AI fairness topics. The model produced fluent and plausible outputs, but the citations it offered consistently favored well-known Western authors and high-impact venues. Table 2 summarizes the patterns observed across four representative prompts. This small test illustrates how even in simple prompts, citation patterns can echo dominant power structures. A reader relying on this output would walk away with a skewed sense of whose work defines the field.

Prompt	**Top Cited Authors**	**Venues**	**Geographic Origin**	**Diversity**
Write a short literature review on fairness in AI.	Barocas, Binns, Dwork	FAccT, NeurIPS	US, UK	Low
Summarize key papers on gender bias in machine learning.	Zhao, Bolukbasi, Kiritchenko	ACL, EMNLP	US, Canada	Moderate
What are key citations on AI ethics and global perspectives?	Floridi, Cowls, Jobin	Nature, AI and Society	UK, Switzerland	Very Low
Generate 5 references on decolonial perspectives in AI.	None (hallucinated or generic)	None	No clear source	Absent

Table 2: Citation behavior in GPT-4 generated responses to literature-related prompts. Cited sources were mostly from well-known Western institutions, with limited or no inclusion of Global South or interdisciplinary perspectives.

This aligns with recent research on hallucinated citations in large language models. Ji et al. [13] find that LLMs frequently generate references that are fabricated but statistically plausible, often echoing the structure and sources dominant in their training data. These hallucinations are not random, they tend to favor well-known authors, major journals, and topics with high publication volume. In effect, even made-up citations can replicate academic power dynamics. Calls for responsible AI use in science have started to address this, but implementation remains limited.

4 Unseen Blind Spots: Why Citation Fairness Remains Ignored in AI Research

Despite growing awareness of AI's social and ethical risks, the specific issue of citation bias in scientific writing tools has received surprisingly little attention. Most fairness research in natural language processing has focused on tasks like sentiment classification, co-reference resolution, or content moderation, areas with clearer demographic benchmarks. But when it comes to academic writing, especially citation behavior, the field still lacks basic metrics, datasets, or evaluation standards. Several benchmark datasets exist for testing gender or racial bias in text generation, like WinoBias (Zhao et al.) [14] or BOLD (Dhamala et al.) [15], but these focus on general language use. They don't cover the epistemic dimension: whose knowledge gets cited, which journals are privileged, or how topic diversity is preserved in scientific communication. Birhane et al. [3] argue that fairness frameworks in AI should include epistemic concerns, not just statistical parity. But even they note that most current tools aren't designed to capture knowledge-level inequality. This gap in attention is not unique to citation systems. Blodgett et al. [16], in their critical survey of fairness in NLP,

point out that most bias research has focused on socially salient categories like race or gender in classification tasks, leaving out epistemic dimensions such as who gets cited or remembered in scientific discourse. This concern is echoed from outside computer science as well. In her work on epistemic injustice, Fricker [17] argues that structural inequality often manifests in credibility judgments, decisions about whose knowledge is taken seriously. When these judgments are automated by LLMs trained on citation rich corpora, they risk reinforcing patterns that already privilege certain voices and traditions.

Meanwhile, developers of LLMs rarely disclose the citation composition of their training data. We don't know how many papers come from high-impact Western journals versus regional or open-access sources. This lack of transparency makes it difficult to evaluate whether the models reproduce citation hierarchies. Without such information, calls for "responsible AI" in science remain abstract. This blind spot is particularly dangerous because AI-generated text can appear neutral and polished, even when it's quietly amplifying structural bias. A paper produced with LLM assistance may look rigorous but rely on a narrow set of citations. And unlike human reviewers, most LLMs can't be held accountable for citation choices, they don't explain or justify them.

If we're serious about epistemic fairness, this is the gap we need to close. We need new benchmarks, datasets, and evaluation methods that treat citation equity as central, not optional, for fair AI in science.

Recognizing the gaps is not enough. In this final section, we outline several concrete steps researchers, developers, and institutions can take to improve citation fairness in the age of AI.

5 Call to Action and Conclusion

To move toward more equitable citation practices in AI-generated academic writing, we need metrics that go beyond counting citations. Table 3 outlines several possible dimensions that could guide future benchmarks and evaluations. These criteria focus on epistemic inclusiveness, not just demographic representation, and reflect a broader concern with whose knowledge systems are being preserved and amplified by AI.

Dimension	**Why It Matters**	**Possible Metrics / Methods**
Author Diversity	Ensures visibility for scholars from underrepresented genders, regions, and institutions.	Geographic spread, gender analysis, institutional affiliation of cited authors.
Journal Diversity	Prevents over-reliance on high-impact or elite journals, which skews epistemic representation.	Distribution across open access, regional, and low-impact journals.
Disciplinary Breadth	Encourages inclusion of interdisciplinary or emerging perspectives often excluded from dominant fields.	Topic modeling, field taxonomy mapping.
Epistemic Style Diversity	Helps avoid reproduction of dominant rhetorical or methodological frameworks only.	Citation sentiment analysis, genre/stylistic profiling.
Citation Novelty	Reduces citation inertia and self-reinforcing loops by promoting less-cited but relevant works.	Citation entropy scores, first-time citation indicators.

Table 3: Suggested dimensions for evaluating citation fairness in AI-generated academic writing. These indicators aim to support epistemic diversity, not just statistical balance.

The rise of LLMs in academic writing is not just a technical development; it is a shift in the way scientific knowledge is created, shared, and remembered. These systems don't just automate text; they automate choices about which voices to amplify and which to leave out. In doing so, they can unintentionally reproduce the same inequalities that have long shaped academic recognition. The problem is not just that citation bias exists, it's that we have few tools to detect or mitigate it when AI systems are involved. Most current evaluations of LLMs say little about epistemic fairness. There are no standard benchmarks to assess whether a model over-represents certain journals, institutions, or regions in its citation suggestions. There's also no agreement on what a "fair" citation profile should even look like in the context of scientific writing. Fixing this will require effort from multiple directions. First, NLP researchers should begin building new benchmarks that measure citation diversity and topical representation in AI-generated academic texts. These benchmarks should be sensitive not just to who is cited, but to what kinds of knowledge are included or ignored. Second, developers of LLMs for academic use should be more transparent about training data, especially the sources of scientific texts and their citation patterns. Third, institutions and publishers need to treat citation equity as part of responsible research practice, particularly when AI tools are involved.

Finally, we need a shift in mindset. Fairness in AI isn't only about demographics or toxicity, it's also about epistemic inclusion: whose work gets seen, cited, and remembered in a world increasingly written by machines. If we ignore that, we risk building a future where scientific diversity is flattened by statistical averages, and knowledge becomes less plural, less global, and less just.

References

[1] Andrew J Peterson. Ai and the problem of knowledge collapse. *AI & Society*, pages 1–21, 2025.

[2] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019.

[3] Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Roni Dotan, and Michael Bao. The values encoded in machine learning research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 173–184. ACM, 2022.

[4] Neven Caplar, Sebastiano Tacchella, and Sune Birrer. Quantitative evaluation of gender bias in astronomical publications from citation counts. *Nature Astronomy*, 1(6):0141, 2017.

[5] Maarten L Wuestman, Jarno Hoekman, and Koen Frenken. The geography of scientific citations. *Research Policy*, 48(7):1771–1780, 2019.

[6] Eitan Frachtenberg and Kyle S McConville. Metrics and methods in the evaluation of prestige bias in peer review: A case study in computer systems conferences. *PLOS ONE*, 17(2):e0264131, 2022.

[7] Maria do Mar Pereira. Rethinking power and positionality in debates about citation: Towards a recognition of complexity and opacity in academic hierarchies. *The Sociological Review*, 2024.

[8] Roberta Sinatra, Dashun Wang, Pierre Deville, Chaoming Song, and Albert-László Barabási. Quantifying the evolution of individual scientific impact. *Science*, 354(6312):aaf5239, 2016.

[9] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.

[10] Langdon Winner. Do artifacts have politics? In *Computer ethics*, pages 177–192. Routledge, 2017.

[11] Liang Huang, Wei Yu, Wei Ma, Wei Zhong, Zhen Feng, Hao Wang, Qiaoqiao Chen, Wei Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.

[12] Wenliang Liang, Yuke Zhang, Ziyang Wu, Hanna Lepp, Wei Ji, Xiaotao Zhao, Hongbo Cao, Shufan Liu, Shijin He, Zhihua Huang, et al. Mapping the increasing use of llms in scientific papers. *arXiv preprint arXiv:2404.01268*, 2024.

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

[14] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. In *arXiv preprint arXiv:1804.06876*, 2018.

[15] Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 862–872. ACM, 2021.

[16] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*, 2020.

[17] Miranda Fricker. Hermeneutical injustice. *Epistemic injustice: Power and the ethics of knowing*, pages 147–175, 2007.