Attributing and situating knowledge cannot be left to language models

Meticulous citation is the marker of well-researched, serious scholarship. Citations do a lot more than attributing credit; they situate claims within the context of existing research and enable scrutiny. When authors cite carelessly, e.g. by referencing famous figures and articles while overlooking original sources, they make two important errors. First, they credit ideas to the wrong person and, second, they reveal a limited understanding of the relevant scholarship. Misattribution disproportionately harms underrepresented voices, whose work has been shown to be consistently more novel than that of established researchers[1]. Research led by women tends to be less cited than comparable research led by men[2 3]. Similarly, research from underrepresented groups, as well as from amateurs and beginners, tends to lead to more breakthroughs and innovation, yet remains less cited than follow-up work from established researchers[4 5 6].

In universities, LLMs are now widely used by both students and researchers as part of their research workflow. Students use generative AI tools primarily to “create and improve educational content”, an unsurprising statistic revealed by Anthropic[7]. Researchers resort to them extensively to draft and edit articles—with new policies by journals and conferences increasingly allowing what is called an ethical and responsible use. The use of LLMs in scholarly writing could turbocharge research by making it faster, more effective, and less stressful, but risks undoing decades of progress in establishing standards for research excellence[8].

In their letter, Earp et al. correctly identify an important issue with LLM use in academia: we cannot identify the ‘provenance’ of generated texts, which they describe as a “systematic breakdown of the chain of academic acknowledgement”[9]. Because LLMs are trained on large-scale text corpora, they cannot provide reliable, auditable attribution for the ideas they incorporate. Developing methods to enable such attribution is expected to be a significant technical challenge. Even if this were possible, we believe that LLMs would likely identify famous authors and easily

---

[1] Hofstra, B., Kulkarni, V. V. , Munoz-Najar Galvez, S., He, B., Jurafsky, D. & McFarland, A. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 9284–9291 (2020).
[2] Budrikis, Z. *Nat. Rev. Phys.* **2**, 346–346 (2020).
[3] Rosa, F., Anastácio, K., Pereira de Jesus, M.V., Veras, H. *Telecommunications Policy* **48**, 5 (2025)
[4] Koffi, M., Pongou, R. & Wantchekon, L. https://www.nber.org/papers/w33150 (2024) doi:10.3386/w33150.
[5] Mims, F. M., III. *Science* **284**, 55–56 (1999).
[6] Kavanagh, K. *Nature* **646**, 269–270 (2025).
[7] Anthropic. https://www.anthropic.com/news/anthropic-education-report-how-university-students-use-claude (2025).
[8] Pearson, H. *Nature* **646**, 788–791 (2025).
[9] Earp, B. D., Yuan, H., Koplin, J. & Porsdam Mann, S. *Nat. Mach. Intell.* **7**, 1889–1890 (2025).

accessible articles at the expense of original ideas. Indeed, research shows that LLMs reflect human citation biases towards highly cited papers[10].

However, we disagree with one of the possible solutions envisioned by Earp et al., which suggests moving away from attribution towards a more collaborative model of scholarship that would treat knowledge as co-produced by humans and machines. This new model would attempt to fix an increasingly broken academic system, where citations are often manipulated to maximise metrics that influence career prospects and reputation, and where attribution is challenging in fast-paced fields with overwhelming volumes of work being produced. Yet, eliminating attribution would not improve academia. On the contrary, it would make it harder for underrepresented voices, both within and outside academia, to be heard. Failing to diligently trace ideas back to their original authors would deepen the erasure of their scholarship and set back their careers.

A more defensible conclusion is that LLMs are not authors and cannot bear accountability for human choices made by AI developers and subsequently by users. Researchers who make use of LLM-generated texts must remain responsible for every claim made, and ensure that ideas are attributed to their original authors so that under-credited work is not further erased. Academia as a whole will benefit from it, preserving thought and voice diversity over the levelling-down that LLMs are introducing.

Author information

Authors and Affiliations

Roxana Radu, Blavatnik School of Government, University of Oxford [roxana.radu@bsg.ox.ac.uk]
Luc Rocher, Oxford Internet Institute, University of Oxford

Corresponding author

Correspondence to Roxana Radu [roxana.radu@bsg.ox.ac.uk]

Ethics declarations

Competing interests

The authors declare no competing interests.

[10] Algaba, A., Mazijn, C., Holst, V., Tori, F., Wenmackers, S. & Ginis, V. In *Findings of the Association for Computational Linguistics: NAACL 2025* 6829–6864 (ACL, 2025).