AI, Access, and Acknowledgement

written by Jeremy Huggett | 09/01/2026



Original photo by Markus Winkler on Unsplash

A number of archaeological/humanities websites are reporting problems with the sheer volume of traffic caused by web crawlers belonging to the various large language model (LLM) generative artificial intelligence systems like ChatGTP, Gemini, CoPilot, and Claude, for instance. Their impact can have the same effect as a Distributed Denial of Service attack, overwhelming the systems. Of late, Knowledge Commons has reported such problems (e.g. Mastodon post), as has Open Context (e.g. Blog post). As Eric Kansa describes it, the AI bots frequently ignore directives designed to protect the site against such activities while Kathleen Fitzpatrick says that dealing with the bots sent out by extractive AI companies threatens the future development of the Knowledge Commons site. Indeed, in the replies to her post, it's clear that the problem is experienced widely – Reclaim (the company hosting this site) report similar problems, while Tom Elliot suggests the Pleiades Gazetteer might have to require logins even for read-only access. Hosting and internet infrastructure companies can do their best to block excessive traffic, but a blanket ban on access – by, for instance, disallowing an AI's user agents via a robots.txt file (assuming they take notice of it) or trying to block them at the firewall – isn't necessarily desirable if you specifically want your information to get wide circulation and use. But if you go down a more permissive route, what then?

A paper in *Nature Machine Intelligence* (Earp et al. 2025) has received a lot of attention in the press and elsewhere of late. In it, they argue that scholarly writing has a provenance problem which a human user prompting the LLM will be largely unaware of:

> … unbeknown to us, the specific substance and flow of ideas in the paragraph closely track those of a now-obscure paper by Smith (1975). Suppose the LLM trained on, but does not mention, Smith's text. And suppose we have never read it or even heard of Smith or her work. So, we don't cite the paper either. Still, our essay now inherits — via nebulous, machine-mediated means — a distinctive insight that Smith developed but for which she

> receives no credit. (Earp et al. 2025, 1889).

As they point out, traditional approaches would encounter Smith's work either directly or through following a citation trail, which should ultimately result in Smith being credited in the resulting text. However, as a result of the growing use of generative AI, scholarship may go unrecognised and past errors may be reproduced, and this provenance problem only worsens as LLMs become more proficient in generating text, as we're already seeing in the parallel development of LLM-generated imagery which is becoming increasingly difficult to distinguish. The solutions that Earp et al. offer range from simply not using LLMs in the first place, to using them to assist with existing text rather than generating its own, or to simply disclose the use of LLMs to act as some kind of warning to readers. They conclude by suggesting that we should perhaps move to a more collaborative form of scholarship involving both humans and machines (Earp et al. 2025, 1890).

John Naughton's commentary on Earp et al. (Naughton 2025) offers some similar approaches to mitigating the problem, not least an emphasis on being appropriately sceptical and requiring verification for anything an LLM produces. On the other hand, David Berry (2025) takes a different approach, arguing that scholarly writing has been influenced by computational tools for many years (word processors, spell- and grammar-checkers, bibliographic management tools, and so on). It's an interesting point: we're aware how using a word processor changes the way we write, for example, yet we don't feel the need to disclose that it or any of its associated tools have been used in our writing. Consequently, Berry asks why LLM use seems to have precipitated a selective anxiety about LLMs, and suggests it is less to do with ethics and more to do with the greater visibility of the way that they mediate thought:

> … academic writing has always been a kind of bricolage, authorship has always been a construction, and … the fantasy of individual academic creation has never corresponded to the actual practices of scholarly production. This would mean that LLMs do not introduce a radical break in how academic knowledge is produced, rather, they reveal and intensify processes that were always operative but remained partially obscured by romantic ideologies of authorship and creativity. (Berry 2025).

So he suggests that the problem is not so much a breakdown in the practice of academic attribution but more a case of the revelation of its already-existing limitations:

> The question is not whether we can return to a world where information had a body and authors had stable identities, but what kind of postdigital scholarship we will build in the world where both have irrevocably flickered into pattern, and where diffusionisation has become the dominant mode of textual production through probabilistic media. (Berry 2025).

Nevertheless, we might reasonably expect that when our data or ideas are reused by others (including computational systems) that they are appropriately acknowledged, as is the expectation in the 'traditional' academic sphere. That traditional expectations are by no means foolproof – we've all had the experience of reading something and thinking "well, I said exactly that X years ago in

paper Y" – is not a reason to not pursue it. Failure to properly cite sources can lead to accusations of plagiarism or worse, and yet that is frequently the case with chatbots who often fail to cite their sources and when challenged, may invent non-existent sources. How then to balance the voracious demand of AI bots for information against anxieties about appropriate acknowledgement through citation?

Eric Kansa recently pointed to Really Simple Licensing (RSL) as a means of making our expectations about the access and use to our information explicit to computational tools. RSL is a new standard for expressing licensing (and compensation) rules about content specifically targeted at AI. It involves providing an XML document containing an RSL statement defining the terms of access and subsequent use. For instance, this blog uses a Creative Commons Attribution-NonCommercial-ShareAlike license which permits sharing and adaptation in non-commercial contexts but requires attribution. However, this was only presented in a human-readable form on every page, whereas RSL presents it in a simple machine-readable form to a visiting bot, informing it of the license conditions governing use. Hence in this case:

```
<rsl xmlns="https://rslstandard.org/rsl">
  <content url="/">
    <license>
      <permits type="usage">all</permits>
      <payment type="attribution">

<standard>https://creativecommons.org/license
s/by-nc-sa/4.0/</standard>
      </payment>
    </license>
  </content>
</rsl>
```

Whether of course it has any effect is another matter, not least because it presumes that an AI bot will address the requirement rather than ignore it, which may be a forlorn hope but it is at least another step in mitigation and it makes me feel better …

References

Berry, D. 2025 'Provenance anxiety: death of the author in the age of large language models',

https://stunlaw.blogspot.com/2025/12/provenance-anxiety-death-of-author-in.html

Earp, B.D., Yuan, H., Koplin, J. and Porsdam Mann, S. (2025) 'LLM use in scholarly writing poses a provenance problem', *Nature Machine Intelligence*, 7(12), pp. 1889–1890. https://doi.org/10.1038/s42256-025-01159-8.

Naughton, J. 2025 'AI has a 'provenance problem'. Here's how to avoid it', *The Observer*, 3rd January 2025
https://observer.co.uk/news/columnists/article/ai-has-a-provenance-problem-heres-how-to-avoid-it