Must peer reviewers be human to assess quality?

LLMs can churn out a plausible report, but when it comes to sound research judgement, people power is still seen as a safer option. **By Jackson Ryan**

Do squirrel surgeons generate more citation impact? The question seems ludicrous, or perhaps the start of a bad joke. But the question, posed by data scientist, Mike Thelwall, was not a joke. It was a test. Thelwall, who works at the University of Sheffield, UK, had been assessing the ability of large language models (LLMs) to evaluate academic papers against the criteria of the research excellence framework (REF), the United Kingdom's national audit of research quality. After giving a custom version of ChatGPT the REF's criteria, he fed 51 of his own research works into the model and was surprised by the chatbot's capability to produce plausible reports. "There's nothing in the reports themselves to say that it's not written by a human expert," he says. "That's an astonishing achievement."

However, the squirrel paper really threw the model. Thelwall had created the paper by taking one of his own rejected manuscripts on whether male surgeons generate more citation impacts than female surgeons, and to make it nonsensical he replaced 'male' with 'squirrel', 'female' with 'human' and any references to gender he switched to 'species' throughout the paper. His ChatGPT model could not determine that 'squirrel surgeons' were not a real thing during evaluation and the chatbot scored the paper highly.

Thelwall also found that the model was not particularly successful at applying a score based on REF guidelines to the 51 papers that were assessed. He concluded that as much as the model could produce authentic-sounding reports, it wasn't capable of evaluating quality.

The rapid rise of generative artificial intelligence (AI) such as ChatGPT and image generators such as DALL-E has led to increasing discussion about where AI might fit into research evaluation. Thelwall's study[1], published in May, is just one piece of a puzzle that academics, research institutions and funders are trying to piece together. It comes as researchers also grapple with the many other ways that AI is affecting science and the developing guidelines that are springing up around its use. These discussions, however, have rarely focused on providing a steer on how AI might be used in assessing research quality. "That is the next frontier," says Gitanjali Yadav, a structural biologist at India's National Institute of Plant Genome Research in New Delhi, and member of the AI working group at the Coalition for Advancing Research Assessment, a global initiative to improve research assessment practice.

Notably, the AI boom also coincides with growing calls to rethink how research outputs are evaluated. Over the past decade, there have been calls to move away from publication-based metrics such as journal impact factors and citation counts, which have shown to be prone to manipulation and bias. Integrating AI into this process at such a time provides an opportunity to incorporate it in new mechanisms for understanding, and measuring, the quality and impact of research. But it also raises important questions about whether AI can fully aid research evaluation, or whether it has the potential to exacerbate issues and even create further problems.

> **"There's no kind of glory or funding associated with peer review. It's just seen as a scientific duty."**

Quality assessments

Research quality is difficult to define, although there is a general consensus that good quality research is underpinned by honesty, rigour, originality and impact. There's a wide variety of mechanisms, each operating at different levels of the research ecosystem, to assess these traits, and myriad ways to do so. The bulk of research-quality assessment happens in the peer-review process, which is, in many cases, the first external quality review performed on a new piece of science. Many journals have been using a suite of AI tools to supplement this process for some time. There's AI to match manuscripts with suitable reviewers, algorithms that detect plagiarism and check for statistical flaws, and other tools aimed at strengthening integrity by catching data manipulation.

More recently, the rise of generative AI has seen a rush of research aimed at exploring how well an LLM might be able to aid peer review — and whether scientists would trust those tools to do so. Some publishers allow AI to assist in manuscript preparation, if adequately disclosed, but do not allow its use in peer review. Even so, there's a growing belief among academics in the ability of these tools, particularly those based on natural language processing and LLMs.

A study published in July this year[2], led by computer science PhD student, Weixin Liang, in the lab of biomedical data scientist, James Zou, at Stanford University in California, assessed the capability of one LLM, GPT-4, to provide feedback on manuscripts. The study asked researchers to upload a manuscript and have it assessed by their AI model. Researchers then completed a survey evaluating the feedback and how it compared with human reviewers. It received 308 responses, with more than half describing the AI-generated reviews as "helpful" or "very helpful". But the study did highlight some problems with that feedback: it was sometimes generic and struggled to provide in-depth critiques.

Zou thinks this doesn't necessarily preclude the use of such tools in certain situations. One particular example he mentions is early-career researchers working on the first draft of a paper. They could upload a draft to a bespoke LLM and receive commentary about deficiencies or errors in their draft. But given the laborious and somewhat repetitive nature of peer review, some academics worry that there could be a tendency to lean on the outputs from a generative AI system capable of delivering reports. "There's no kind of glory or funding associated with peer review. It's just seen as a scientific duty," says Elizabeth Gadd, head of research culture and assessment at Loughborough University, UK. There is already evidence that peer reviewers are using ChatGPT and other chatbots to some extent, despite the rules put in place by some journal publishers.

NEIL WEBB

Thelwall believes there's more that AI could do in helping peer reviewers to evaluate research quality, but there is reason to move slowly. "We just need lots of testing," he says. "And not just technical testing, but also pragmatic testing, where we gain confidence that if we provide the AI to the reviewers, for example, that they won't abuse it."

Yadav sees great benefit in AI as a time-saving tool and has been working with it to help rapidly assess wildlife imagery from field-based cameras in India, but she sees peer review as too important to the scientific community to hand over to the bots. "I'm personally absolutely against peer review being done by AI," she says.

Quality savings

One of the most discussed benefits of using AI is the idea that it could free up time. This is particularly apparent in institutional and national systems of evaluating research – some of which have incorporated AI. For instance, one funder in Australia, the National Health and Medical Research Council (NHMRC), already uses AI through "a hybrid model combining machine learning and mathematical optimisation techniques" to identify suitable human peer reviewers to judge grant proposals. The system helps to remove one of the administrative bottlenecks in the evaluation process, but it's where the AI use ends. An NHMRC spokesperson says the agency "does not use artificial intelligence, in any form, to directly assist with research quality evaluation" itself.

Even using AI for such administrative support could be a major resource saving, however, especially for large national assessments such as the REF. Thelwall says the exercise is known for its incredible drain on researchers' time. More than 1,000 academics help to assess research quality in the REF and it takes them about half a year to get it done.

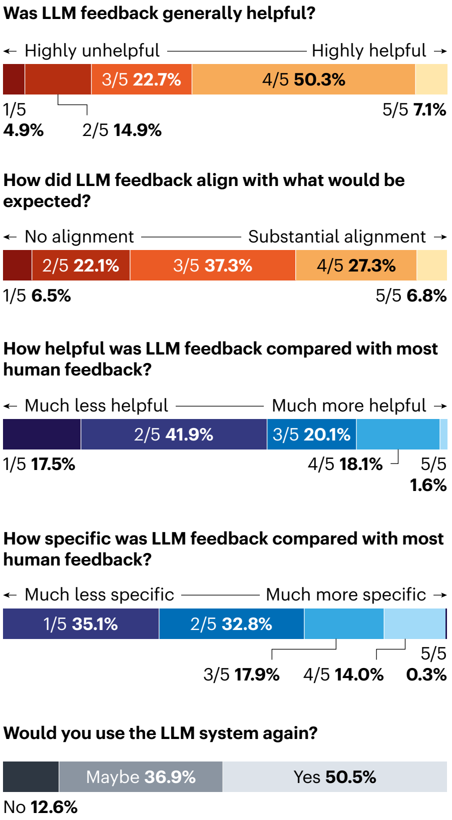
"If we can automate evaluations", says Thelwall, then "it would be a massive productivity boost". And there's potential for huge savings: the most recent REF, in 2021, was estimated to have cost around £471 million (US$618 million).

Similarly, New Zealand's assessment of researchers, the Performance Based Research Fund, has previously been described by Tim Fowler, chief executive of the government's Tertiary Education Commission, as a "backbreaking" exercise. In it, academics submit portfolios for assessment, placing an extreme burden on them and institutions. In April, the government scrapped it and a working group has been charged with delivering a new plan by February 2025.

These examples suggest AI's major potential to create more efficiency, at least for large, bureaucratic, assessment systems and processes. At the same time, the technology is developing as perspectives on what constitutes research quality are evolving and becoming more nuanced. "How you might have defined research quality in the early twentieth century is not how you define it now," says Marnie Hughes-Warrington, deputy vice-chancellor of research and enterprise at the University of South Australia in Adelaide. Hughes-Warrington is a member of the Excellence in Research Australia transition group, which is considering the future of the country's assessment exercise after a review in 2021 found that it placed a significant burden on universities. She says the research community is increasingly recognizing the need to assess more "non-traditional research outputs" – such as policy documents, creative works, exhibitions – and then beyond to social and economic impacts.

MIXED REVIEWS

Around 300 researchers who were asked to rate a large language model's (LLM) ability to provide feedback on manuscripts found its comments to be helpful and in alignment with the type of comments they would expect. Half also said they would use the system again. But compared with human reviewers, most researchers thought the AI feedback was less specific and tended to be less helpful.



Was LLM feedback generally helpful?

How did LLM feedback align with what would be expected?

How helpful was LLM feedback compared with most human feedback?

How specific was LLM feedback compared with most human feedback?

Would you use the LLM system again?

LIANG, W. *ET AL. NEJM AI* HTTPS://DOI.ORG/10.1056/AIOA2400196 (2024).

As the conversations are happening alongside the AI boom, it makes sense that new tools could fit into revised methods of research-quality evaluation. For instance, Hughes-Warrington points to how AI is already being used to detect image manipulation in journals or to synthesize data from systems used to uniquely identify researchers and documents. Applying these kinds of methods would be consistent with the mission of institutions such as universities and national bodies. "Why wouldn't organizations, driven by curiosity and research, implement new ways of doing things?" she says.

However, Hughes-Warrington also highlights where incorporating AI will meet resistance. There's privacy, copyright and data-security concerns to acknowledge, inherent biases in the tools to overcome and a need to consider the context in which research assessments take place, such as how impacts will differ across disciplines, institutions and countries.

Gadd isn't against incorporating AI and says she is noticing it appear more often in discussions around research quality. But she warns that researchers are already one of the most assessed professions in the world. "My own general view on this is that we assess too much," she said. "Are we looking at using AI to solve a problem that's of our own making?"

Having seen how bibliometrics-based assessments can damage the sector, with metrics such as journal impact factors misused as a substitute for quality and shown to hinder early-career researchers and diversity, Gadd is concerned about how AI might be implemented, especially if models are trained on these same metrics. She also says decisions involving allocation of promotions, funding or other rewards will always need human involvement to a far greater extent. "You have to be very cautious", she says, about shifting to technology "to make decisions which are going to affect lives".

Gadd has worked extensively in developing SCOPE, a framework for responsible research evaluation by the International Network of Research Management Societies, a global organization that brings research management societies together to coordinate activities and share knowledge in the field. She says one of the key principles of the scheme is to "evaluate only where necessary" and, in that perhaps, there is a lesson for how we should think about incorporating AI. "If we evaluated less, we could do it to a higher standard," she says. "Maybe" AI can support that process, but a "lot of the arguments and worries we're having about AI, we had about bibliometrics."

Jackson Ryan is a freelance journalist in Sydney, Australia.

1. Thelwall, M. *J. Data Inform. Sci.* **9**, 1–21 (2024).
2. Liang, W. *et al. NEJM AI* https://doi.org/10.1056/AIoa2400196 (2024).