https://doi.org/10.51408/issi2025_156

Responsible Uses of Large Language Models for Research Evaluation

Mike Thelwall

m.a.thelwall@sheffield.ac.uk
Information School, University of Sheffield (UK)

Abstract

Although research evaluators and scientometricians have promoted the message of responsible bibliometrics through initiatives like the Leiden Manifesto, these do not mention Large Language Models (LLMs). LLMs can now make useful quality predictions for journal articles, giving values that correlate more strongly with expert judgements than do citation-based indicators in most fields. This has created the possibility that they could supplement or even replace citation-based indicators for some applications. As tested so far, LLMs predict the quality rating that a human expert would give a paper. They do this by reading the quality level descriptions and then processing the article title and abstract. This raises multiple new issues in comparison to the Leiden Manifesto. First, authors might try to trick LLMs into giving high scores by crafting LLM-friendly abstracts. Second, LLM models incorporate billions of parameters, so their scores are opaque. Third, it is not clear how LLMs work in terms of the main influences on their scores, so their biases are unknown. Fourth, whilst citations reflect tangible and permanent contributions to the scientific record, albeit of variable value, LLM-based predictions do not clearly link to progress. Fifth, LLM scores are ephemeral in the sense that newer LLMs may give substantially different scores and rankings.

Introduction

Research evaluation is often used to support decision making. For example, job applicants may be judged on the quality of their work, departmental funding might be dependent on positive research quality or volume evaluations, and national policy may be informed by estimates of the areas in which the country appears strong relative to its competitors. In these cases, there are winners and losers, assuming that there are finite resources to allocate. Thus, it is not only important to ensure that the research evaluations are as accurate as possible, but also that they are not biased in a way that would undermine the system. These two considerations do not always fully align: if the research evaluation method that is the most accurate overall also has a substantial bias against a particular group (e.g., women, ethnic minorities, applied researchers), then it might not be acceptable for reasons of social equity or national policy.

A research evaluation approach might also be ruled irresponsible if it generates perverse incentives. Whenever people are evaluated and know the evaluation method, it is natural for some to target the method rather than the underlying goal, potentially generating unwanted outcomes. For example, if academics are evaluated on the number of articles they produce then they might divert some of their effort into publishing smaller and possibly weaker articles at the expense of books, chapters, conference papers, and long articles (Aagaard, 2015).

A third issue is transparency: the ability of those evaluated, or affected by an evaluation, to see the details of the mechanism used to evaluate them. This may give confidence in the evaluation system and may improve it if errors can be identified and corrected. In practice, transparency is always partial. The most transparent system might be citation-based indicators from OpenAlex since it publishes the source code of all the algorithms it uses (Priem et al., 2022). This is still not full transparency because its citation counts are based on citations made by millions of individual scientists behind closed doors. A deliberate lack of transparency is also common in research: authors are rarely told the identities of the reviewers rejecting their paper or giving a low score to their grant (i.e., single/double blind peer review), and some decisions are made without any explanation or rationale.

In the light of these considerations, it seems reasonable to suggest that research evaluations ought to be as responsible as possible, in the sense of minimising the above risks as far as is practical in the context of the goals and resources of the evaluation. It also seems like good practice to be honest about the extent to which any problems remain. The rest of this paper briefly summarises some responsible research evaluation initiatives for bibliometrics and then focuses on the considerations that are relevant to the use of large language models to support research evaluation.

Responsible bibliometrics

Perhaps the most well-known responsible bibliometrics initiative is the Leiden Manifesto (Hicks et al., 2025). Its ten principles are:

1. Quantitative evaluation should support qualitative, expert assessment.
2. Measure performance against the research missions of the institution, group, or researcher.
3. Protect excellence in locally relevant research.
4. Keep data collection and analytical processes open, transparent, and simple.
5. Allow those evaluated to verify data and analysis.
6. Account for variation by field in publication and citation practices.
7. Base assessment of individual researchers on a qualitative judgement of their portfolio.
8. Avoid misplaced concreteness and false precision.
9. Recognize the systemic effects of assessment and indicators.
10. Scrutinize indicators regularly and update them (Hicks et al., 2025).

Overall, the Leiden Manifesto goal is to reduce the chance that bibliometrics are used unwisely for research evaluation. The Metric Tide (Wilsdon et al., 2013) is similar but more UK focused.

There are other prominent initiatives against inappropriate uses of specific types of indicators as part of a wider movement for assessment reform (Rushforth, 2025; Rushforth & Hammarfelt, 2023). The San Francisco Declaration on Research Assessment (DORA; sfdora.org) campaigns against overuse of journal-based

indicators in the belief that research evaluation should focus on articles rather than publication venues and that focusing too much on journals creates a perverse incentive that is unhealthy to the diversity of scientific publishing. This follows many years of criticisms of article-level citation-based indicators and journal impact factors (e.g., MacRoberts & MacRoberts, 2018; Rushforth & de Rijcke, 2015; Seglen, 1998).

In parallel, More Than Our Rank (inorms.net/more-than-our-rank) campaigns against reliance on league tables of universities. Focusing on league tables can cause perverse incentives, such as hiring academics for their citation rates or prizes rather than their ability to support the university goals (if different). These league tables usually either rely on citation rates or have them as an important component but the other methods used are also flawed. For example, reputational surveys favour older and larger institutions because more academics will know them, giving them a larger potential voter base (Gadd, 2020; Vernon et al., 2018).

As these examples show, specific problems with bibliometrics and related research evaluation methods have given rise to initiatives to combat them. With the rise of Artificial Intelligence (AI) support for research evaluation, potential new problems must also be considered.

Research evaluation applications of LLMs

There have been many attempts to introduce AI in the form of traditional machine learning into research evaluation, such as to predict long term citation rates for recently published articles (Ma et al., 2021), but they do not seem to have led to any practical applications. The situation seems set to change with the rise of Large Language Models (LLMs), which have some capability to follow human instructions for text processing tasks (Ouyang et al., 2022) and perform well in many cases (Kocoń et al., 2023). In this context, early evidence suggests that they have a technical capability to challenge bibliometrics as the most accurate scientific research quality indicator. Specifically, small-scale studies have shown that research quality judgments by ChatGPT for submitted or published articles correlate positively with private human judgements or scores (Saad et al., 2024; Thelwall, 2024) and in some cases for public scores (Zhou et al., 2024; Thelwall & Yaghi, 2025). In addition, a large-scale study has suggested that ChatGPT quality predictions may correlate more strongly than citation-based indicators with research quality scores for most academic fields (Thelwall & Yaghi, 2024; Thelwall et al., 2025; Thelwall, 2025). Since this accuracy creates the possibility that LLMs may complement or replace citation-based indicators in the future, it is important to consider how this might impact on considerations for responsible indicators.

Possible Applications of LLM research quality scores

In theory, LLMs could be used for most evaluation roles that citation-based indicators currently fill. The main current exception is that some citation indicators are network-based, such as evidence of the countries in which a nation's or journal's citations mainly originate (Schubert & Glänzel, 2006; Zhang et al., 2009). This section discusses some likely research evaluation applications of LLMs.

Support for article-level expert quality ratings

Individual articles sometimes need to be assessed or scored for quality for job-related reasons (appointments, tenure, promotion), impacting academic careers. Currently these evaluations might be formal (e.g., asking experts to read and score articles) or informal (e.g., forming a quick impression of a candidate's research strengths by browsing their CV). Heuristics seem likely to be used for quick informal evaluations and those made by people that are not experts on the candidate's topics. These might typically include journal reputation, journal citation rates, and article citation counts. Overinterpreting the results is a common cause of concern (Rushforth & De Rijcke, 2024). LLMs could be used in a similar way, in theory. In practice, it seems unlikely that LLMs would often be used well in this role since they need some knowledge to set up and their scores need to be rescaled from multiple submissions to be meaningful (Thelwall, 2024). Thus, LLMs are currently more of a threat than a benefit in this role, until a system exists that would generate meaningful score predictions (e.g., scaled to align with human judgement).

LLMs might currently be most useful for large scale formal evaluations like the UK Research Excellence Framework (REF), which individually scores over 100,000 journal articles and uses citation-based indicators in a minor role for some health and physical sciences fields and economics. The citation-based indicators are carefully selected and curated, and the same could be achieved for LLMs scores. They might also be useful for a wider range of fields than bibliometrics, including some where they had a stronger correlation with expert judgements than do citation rates (Yaghi & Thelwall, 2024).

Departmental-level evaluations

In some situations, departments are evaluated as a whole, either individually by benchmarking them against other similar departments or as part of a national evaluation of all departments of a given type. Here, it seems plausible that average LLM scores could be calculated as an additional indicator to citation-based indicators. It would be interesting to see if this helped any department type. Again, LLMs might be used across a wider range of fields than citations currently are.

National and international comparisons

In theory, citation-based bibliometric analyses of national strengths and weaknesses, as included in periodic reports by or for governments (e.g., Science, Research and Innovation Performance of the EU) could be supplemented by a section on LLM scores, potentially expanding indicator coverage beyond fields for which citation-base indicators have the most value.

JIFs

Average LLM scores for articles published by a journal can be calculated as an alternative to the average citation rates of the Journal Impact Factor (JIF) and similar formulae. The results from the two approaches correlate positively and moderately or strongly, depending on the field. Moreover, the LLM version may be fairer to

journals that attract relatively few citations because the citing journals are not included in a citation index (Thelwall & Kousha, 2025).
An advantage of LLM-based journal quality indicators is that they could be based on the most recent year of published articles, rather than older articles, as currently used for all well-known citation-based journal impact indicators. This would make the results more current. A potential disadvantage is that if LLM-based journal ranking becomes common then publishers and editors may attempt to at least partly target the journal's formatting requirements or style guidelines towards LLM-friendly elements. It is not clear what this would entail.

Threats to responsible uses of LLMs

This section discusses three of the Leiden Manifesto's most relevant aspects for LLMs.

Perverse incentives

Since perverse incentives occur by people targeting indicators rather than the underlying goal, the logical LLM perverse incentive is for authors to craft articles for high LLM scores rather than for communicating their research accurately and clearly for a human audience. This could be achieved by entering an article into an LLM and asking it to make suggestions to make it more likely to achieve a high score. This might involve exaggerating the importance of the findings to make the research more like a press release.
Whilst it seems likely that authors would attempt to do this, wasting their time on an unproductive activity, it is not clear that it would work to any great extent. Articles go through peer review and this guards against unsupported claims, so authors enlisting LLM help might find their work more likely to be rejected. Moreover, there are many LLMs, they have different strengths, and they evolve over time so it is not clear that crafting an LLM-friendly article would work even if it passed peer review. In addition, if the practice was suspected then evaluators might try to detect and penalise LLM-supported articles.
Thus, overall, it seems reasonably likely that the main perverse incentive is for authors to waste time on creating LLM-friendly work rather than that these attempts would succeed and lower the accuracy of LLM-based evaluations.

Transparency: Opaque LLM scores

LLMs have arguably the same transparency issues as peer review. In the same way that we can't see the processes going on inside a reviewer's brain when they cogitate over what they have read and experienced, turning their knowledge into a score/judgement and report, we also can't follow the numerous weights (typically above 7 billion even for the smallest model) within an LLM leading to its score and justification. In theory, an LLM could have more transparent inputs than a human reviewer in the sense that it could be trained on a known corpus of work (e.g., everything in OpenAlex), and LLM algorithms are certainly more understandable than human brains, at least in their overall architecture. These seem to be minor differences, however, given the overall complexity of even the smallest LLM. In

contrast, bibliometrics seem to be more transparent. For example, citation-based indicators from OpenAlex are relatively transparent, as argued above. Here the main opaqueness, the citing author decisions, is perhaps less important because each decision is relatively minor if many citations are counted for an evaluation.

Biases: Unknown LLM score influences and biases

Since LLM evaluations are relatively new, little is known about their biases. In contrast, some bibliometrics have been shown to have gender biases (e.g., career citations) and most have international biases, and there may also be institutional, reputational and interdisciplinary disparities (e.g., Paris et al., 1998; Schisterman et al., 2017). For ChatGPT-based evaluations it is known that some fields get substantially higher average scores than others (Thelwall & Yaghi, 2024), but little else is known about any other biases.

Research into AI biases in other contexts has shown that apparently objective mathematical algorithms can be biased if they are fed with unbalanced data or misleading assumptions by their engineers. They can also generate new biases as an unintended side effect of their data and algorithm (Akter et al., 2022; Baeza-Yates, 2016). Thus, it is reasonable to expect that LLMs will have learned biases from their inputs and may also have generated new ones. The extent and nature of these is not known, however. It is therefore an important due diligence step for researchers to test LLM scores for the most likely and worrying types of disparity.

New LLMs Irresponsibility Dimensions?

As shown above, responsible uses of LLMs should consider the same issues as for bibliometrics. There are additional considerations that do not apply to citation-based indicators, and some are discussed here.

Ephemerality and variation between LLMs

An important difference between citations and LLMs currently seems to be that citations are tangible and verifiable, whereas LLM judgements are not. In particular, an author judged to have 20 excellent papers by one LLM might next year be judged to have only five by a different LLM or an improved version of the same one. Whilst this perhaps mirrors the peer opinion situation in the sense that a person's work might go out of fashion, it must be demoralising to know that research achievements can disappear suddenly due to an algorithm change. This might reduce confidence in the research evaluation system, if used for individual academics.

There are at least three ways to address this issue. First, research into the stability of LLM scores might give reassurance that wholescale score shifts, as hypothesised above, are unlikely. Second, only aggregate scores across multiple articles might be used for evaluations, reducing the impact of changes for individual articles. Third, long term evaluation processes might build in stability, such as by fixing a score at a given point in time or altering scores primarily by adding new evaluations rather than replacing old evaluations.

Alignment of prompts with evaluation goals

Unlike citations, LLM prompts might be tailored to the goals of a research evaluation. For example, if the goal is the generic one of assessing research quality, then the prompt might ask the LLM to assess an article for the three core dimensions of rigour, significance and originality (Langfeldt et al., 2020), perhaps tailoring the definitions of these to a local context or with local examples. Responsible evaluators would have the option to tailor their prompts to more specific goals, however, such as value to the national economy or support for United Nations Development Goals. Of course, tailoring the prompts to a particular goal does not mean that the LLM will be capable of responding appropriately.

Lack of connection to research progress

Another dimension of uncertainty for LLM scores is that they do not have a direct theorised connection to research progress. For citations, Merton's (1973) theory posits that citations are scholarly acknowledgements of prior work that has aided the creation of new research. This is an oversimplification since the selection of work to cite is subjective with influential prior work often remaining uncited (e.g., obliteration by incorporation: McCain, 2011) and work without influence being cited (e.g., for background context). Nevertheless, it is still possible to claim that in many fields some citations reflect influence, and the rest are noise, with the latter tending to disappear at a sufficiently high level of aggregation (van Raan, 1998). There does not seem to be a way to mitigate the lack of a tangible connection to research progress for LLM evaluations, although it is the same as for expert opinions.

Cost

The relative costs of LLMs and bibliometric indicators are not yet clear. If wide uptake is to be achieved, LLM scores might need to be offered by citation index providers. These would be able to share the costs of the LLM queries or processing across all users. Citation-based indicators currently (March 2025) have two advantages: there are no providers of LLM-based academic scores and OpenAlex is a free source of citation-based indicators. Of course, the cost of LLM scores includes the personnel costs associated with the skills needed to generate the scores as well as the computing costs.

Summary

As argued above, issues relevant to the responsible use of LLM-based quality scores are partly the same as for bibliometrics and partly different, with some new considerations. Returning to the Leiden Manifesto (Hicks et al., 2025), the LLM adjustments can be summarised as follows.

1. *Quantitative evaluation should support qualitative, expert assessment*. This is the same for LLMs, even though they mimic human peer review more than do citations.
2. *Measure performance against the research missions of the institution, group, or researcher*. The same for LLMs.

3. *Protect excellence in locally relevant research*. The same for LLMs. They may have more capacity to do this than citation-based indicators since LLM prompts could be explicitly tailored to local goals, needs and concepts of research quality.
4. *Keep data collection and analytical processes open, transparent, and simple*. Current major LLMs fail this, as discussed above, with the partial exception that their algorithm architectures are known, a minor advantage over human brains.
5. *Allow those evaluated to verify data and analysis*. Current LLMs fail this because they do not publish their data sources, except that those analysed could replicate the actions of those obtaining the scores, if they publish their prompts and the identity of the LLM used. Because of the random parameters used in LLMs, they will not get the same results and might get substantially different results occasionally. This issue could be addressed by evaluators only using offline LLMs and publishing the random seed values, but this seems like a minor point.
6. *Account for variation by field in publication and citation practices*. Users of LLMs should consider variations between fields in the average LLM scores.
7. *Base assessment of individual researchers on a qualitative judgement of their portfolio*. This is the same for LLMs.
8. *Avoid misplaced concreteness and false precision*. This is the same for LLMs.
9. *Recognize the systemic effects of assessment and indicators*. This is also important for LLMs although the issues are different, as discussed above.
10. *Scrutinize indicators regularly and update them*. This seems likely to happen naturally for LLMs as new versions appear and existing ones evolve. New prompts should also be tested especially to align to local needs.

To these ten points, four additional suggestions could be incorporated for LLM-based scores.

11. *Design prompts to align with the research evaluation goals*.
12. *Consider the ephemerality of scores and differences between LLMs*.
13. *Consider the costs of generating LLM scores relative to bibliometric alternatives.*
14. *Accept that LLM scores are not direct evidence of contributions to science*.

References

Aagaard, K. (2015). How incentives trickle down: Local use of a national bibliometric indicator system. *Science and Public Policy*, *42*(5), 725-737.

Akter, S., Dwivedi, Y. K., Sajib, S., Biswas, K., Bandara, R. J., & Michael, K. (2022). Algorithmic bias in machine learning-based marketing models. Journal of Business Research, 144, 201-216.

Baeza-Yates, R. (2016). Data and algorithmic bias in the web. In Proceedings of the 8th ACM Conference on Web Science. https://doi.org/10.1145/2908131.2908135

Gadd, E. (2020). University rankings need a rethink. Nature, 587(7835), 523-524.

Hicks, D., Wouters, P., Waltman, L., De Rijcke, S., & Rafols, I. (2015). Bibliometrics: the Leiden Manifesto for research metrics. Nature, 520(7548), 429-431.

Kocoń, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydło, D., Baran, J., & Kazienko, P. (2023). ChatGPT: Jack of all trades, master of none. Information Fusion, 99, 101861.

Langfeldt, L., Nedeva, M., Sörlin, S., & Thomas, D. A. (2020). Co-existing notions of research quality: A framework to study context-specific understandings of good research. Minerva, 58(1), 115-137.

Ma, A., Liu, Y., Xu, X., & Dong, T. (2021). A deep-learning based citation count prediction model with paper metadata semantic features. Scientometrics, 126(8), 6803-6823.

MacRoberts, M. H., & MacRoberts, B. R. (2018). The mismeasure of science: Citation analysis. Journal of the Association for Information Science and Technology, 69(3), 474-482.

McCain, K. W. (2011). Eponymy and obliteration by incorporation: The case of the "Nash Equilibrium". Journal of the American Society for Information Science and Technology, 62(7), 1412-1424.

Merton, R. K. (1973). The sociology of science: Theoretical and empirical investigations. University of Chicago Press.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730-27744.

Paris, G., De Leo, G., Menozzi, P., & Gatto, M. (1998). Region-based citation bias in science. Nature, 396(6708), 210-210.

Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. arXiv preprint arXiv:2205.01833.

Rushforth, A. (2025). Research Assessment Reform as Collective Action Problem: Contested Framings of Research System Transformation. Minerva, 1-21.

Rushforth, A., & de Rijcke, S. (2015). Accounting for impact? The journal impact factor and the making of biomedical research in the Netherlands. Minerva, 53(2), 117-139.

Rushforth, A., & De Rijcke, S. (2024). Practicing responsible research assessment: Qualitative study of faculty hiring, promotion, and tenure assessments in the United States. Research Evaluation, 33, rvae007.

Rushforth, A., & Hammarfelt, B. (2023). The rise of responsible metrics as a professional reform movement: A collective action frames account. Quantitative Science Studies, 4(4), 879-897.

Saad, A., Jenko, N., Ariyaratne, S., Birch, N., Iyengar, K. P., Davies, A. M., Vaishya, R., & Botchu, R. (2024). Exploring the potential of ChatGPT in the peer review process: An observational study. Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 18(2), 102946. https://doi.org/10.1016/j.dsx.2024.102946

Schisterman, E. F., Swanson, C. W., Lu, Y. L., & Mumford, S. L. (2017). The changing face of epidemiology: gender disparities in citations? Epidemiology, 28(2), 159-168.

Schubert, A., & Glänzel, W. (2006). Cross-national preference in co-authorship, references and citations. Scientometrics, 69, 409-428.

Seglen, P. O. (1998). Citation rates and journal impact factors are not suitable for evaluation of research. Acta Orthopaedica Scandinavica, 69(3), 224-229.

Thelwall, M., & Yaghi, A. (2024). In which fields can ChatGPT detect journal article quality? An evaluation of REF2021 results. arXiv preprint arXiv:2409.16695.

Thelwall, M. & Yaghi, A. (2025). Evaluating the predictive capacity of ChatGPT for academic peer review outcomes across multiple platforms. Scientometrics, to appear.

Thelwall, M., Jiang, X., & Bath, P. (2025). Estimating the quality of published medical research with ChatGPT. Information Processing & Management, 62(4), 104123. https://doi.org/10.1016/j.ipm.2025.104123

Thelwall, M., & Kousha, K. (2025). Journal Quality Factors from ChatGPT: More meaningful than Impact Factors? Journal of Data and Information Science. https://doi.org/10.2478/jdis-2025-0016

Thelwall, M. (2024). Can ChatGPT evaluate research quality? Journal of Data and Information Science, 9(2), 1–21. https://doi.org/10.2478/jdis-2024-0013

Thelwall, M. (2025). In which fields do ChatGPT 4o scores align better than citations with research quality? *arXiv preprint arXiv:2504.04464.*

van Raan, A. F. (1998). In matters of quantitative studies of science, the fault of theorists is offering too little and asking too much. Scientometrics, 43, 129-139.

Vernon, M. M., Balas, E. A., & Momani, S. (2018). Are university rankings useful to improve research? A systematic review. PloS One, 13(3), e0193762.

Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., Hill, S., & Johnson, B. (2015). The metric tide. Report of the independent review of the role of metrics in research assessment and management. https://www.ukri.org/wp-content/uploads/2021/12/RE-151221-TheMetricTideFullReportLitReview.pdf

Zhang, L., Glänzel, W., & Liang, L. (2009). Tracing the role of individual journals in a cross-citation network based on different indicators. Scientometrics, 81(3), 821-838.

Zhou, R., Chen, L., & Yu, K. (2024). Is LLM a Reliable Reviewer? A Comprehensive Evaluation of LLM on Automatic Paper Reviewing Tasks. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024) (pp. 9340-9351).