
Article:
Thelwall, M. (2026) Large language models and responsible research evaluation: an extension of the Leiden Manifesto. Scientometrics. ISSN: 0138-9130

https://doi.org/10.1007/s11192-026-05552-x

Scientometrics
https://doi.org/10.1007/s11192-026-05552-x

Large language models and responsible research evaluation: an extension of the Leiden Manifesto

Mike Thelwall[1]

Received: 4 August 2025 / Accepted: 13 January 2026

Abstract
Research evaluators and scientometricians sometimes promote the message of responsible bibliometrics through initiatives like the Leiden Manifesto, but these do not mention Large Language Models (LLMs). Since there is evidence that LLMs can make quality predictions for journal articles that correlate more strongly with expert judgements than do citation-based indicators in most fields, they may start to supplement or replace citation-based indicators for some applications. This paper compares responsible evaluation principles from the Leiden Manifesto with those necessary for LLMs, finding them all to be still relevant. It also includes a discussion of how they apply to human expert review and the differences between the three approaches. For example, transparency is inherently weak for LLMs, since their decision-making processes are too complex to fully understand (similarly for experts). Conversely, LLMs may be able to address some issues that citation-based indicators can't, such as by adapting prompts to the goals of an evaluation. Other issues impact differently. For example, LLM evaluations might encourage authors to customize articles to persuade LLMs, whereas bibliometric evaluations might encourage authors to solicit citations. Finally, two additions to the Leiden Manifesto targeting LLM-supported research evaluations are proposed. First, the cost/benefit tradeoff should be considered when deciding which approach to use. In resource-poor contexts or for minor evaluations, it may be irresponsible to implement otherwise fully responsible solutions. Second, LLM evaluations need to comply with national copyright law when processing academic texts: the ability to access a text does not always entail permission to automatically process it.

Keywords Responsible research evaluation · Large language models · Leiden manifesto

Introduction

Research evaluation is often used to support decision-making: 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 is stronger than its competitors. Depending on the scale

✉ Mike Thelwall
m.a.thelwall@sheffield.ac.uk

[1] School of Information, Journalism and Communication, University of Sheffield, Sheffield, UK

of the evaluation and the choices made, the approach may be primarily qualitative in the form of expert review, primarily quantitative, such as with scientometric approaches, or a mix of the two (Moed, 2005; Wilsdon et al., 2015).

Assuming that there are finite resources to allocate, research evaluations always create winners and losers, irrespective of the approach used. 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 goals, including fairness (Dahler-Larsen, 2011). These two considerations are not the same: inaccuracy increases the chance that an individual evaluated entity (e.g., researcher, department, research field) has an inaccurate evaluation, whereas bias entails a group of evaluated entities tending to be given a score that is too high (or too low) (Wilsdon et al., 2015). The two considerations of bias and inaccuracy also 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 be unacceptable for reasons of social equity or national policy. As discussed below, many initiatives for responsible research evaluation centre on bias and inaccuracy and some also foreground transparency as a way for those evaluated to protect themselves against inaccuracy and bias.

It seems uncontroversial to suggest that research evaluations ought to be as responsible as possible, in the sense of minimising inaccuracy and bias and maximising transparency as far as is practical in the context of the goals and available resources (Hicks et al., 2015; Wilsdon et al., 2015). It also seems like good practice to be honest about the extent to which any problems remain. Whilst these issues have been extensively discussed for bibliometrics, Large Language Model (LLM)-based quality scores potentially provide a more useful overall level of quantitative evidence, in the sense of correlating more strongly with expert scores in most fields (Thelwall, 2025a). It is therefore important to revisit bibliometric discussions of responsible bibliometrics to help ensure that applications of LLM scores are responsible from the start. This article first discusses ways in which LLMs can support research evaluations. It then introduces some high-profile responsible research evaluation initiatives for bibliometrics and focuses on the considerations that are relevant to the use of LLMs to support research evaluation. It uses the Leiden Manifesto for bibliometrics as a framework to discuss the main commonalities and differences between LLMs, bibliometrics, and expert review for responsible research evaluations. Finally, it introduces an extension of the Leiden Manifesto for LLMs.

LLM research quality score applications

There have been many attempts to introduce Artificial Intelligence (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). Nevertheless, they do not seem to have led to any practical applications. The situation seems to be changing with the rise of 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). The first research evaluation context that typical academics may meet LLMs is peer review, whether translating their own review or receiving a review from a reviewer that has broken confidentiality and trust by using an LLM to write the review from the article (e.g., Liang et al., 2024).

An LLM is an AI system that has been trained to recognise patterns in text at a deep level (in a multi-layer network) by ingesting a large body of documents using the

"transformer" information processing architecture. The best-known examples of LLMs are the generative type, which can leverage knowledge of text patterns to write good quality text, often in many languages. The well-known LLMs are also given an extra training stage to help them learn how to respond to human requests. This functionality is typically exposed through a chat interface, such as chatgpt.com or deepseek.com or, for direct access by computer programs, through an Applications Programming Interface (API). In addition, "open weights" LLMs can be downloaded as huge files and run locally, which is useful when secure processing is needed.

An important practical aspect of LLMs is the need to curate prompts to elicit the most useful information from them. Whilst they seem to be good at responding appropriately to any reasonably expressed request, prompt optimisation can improve their results (e.g., Wang et al., 2025).

Early evidence suggests that LLMs 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 (Thelwall & Yaghi, 2025b; Wu et al., 2025; Zhou et al., 2024). In addition, a large-scale study has suggested that ChatGPT quality predictions may correlate to a greater extent than citation-based indicators with research quality scores for most academic fields (Thelwall & Yaghi, 2025a; Thelwall et al., 2025; Thelwall, 2025a). 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.

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). No LLM approach so far has generated network type information, although LLMs are networks and might use similar types of networked connected information in their evaluations. 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, in the sense of having too much confidence in the conclusions drawn from this data (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 are not individually meaningful but are only useful when they are transformed into ranks within a set of documents (Thelwall, 2024). Thus, LLMs are currently more of a threat (from evaluators taking individual scores seriously) than a benefit in this role, until a system exists that would generate meaningful score predictions (e.g., scaled to align with human judgement or ranked).

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, 2025a).

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. For example, France uses self-evaluations with follow up visits by experts (Hcéres, 2025). For whole department evaluations, it seems plausible that average LLM scores could be calculated as an additional indicator to citation-based indicators, if used. 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: EC, 2024) could be supplemented by a section on LLM scores, potentially expanding indicator coverage beyond fields for which citation-based 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. Of course, many or all problematic uses of JIFs in research evaluation would also be problematic for LLM variants.

Responsible bibliometrics initiatives

This section focuses only on responsible bibliometrics because general discussions of responsible AI are mostly not directly relevant (Zhang et al., 2024). Considerations of copyright, security and privacy for AI are important, however.

The most well-known general responsible bibliometrics initiative is the Leiden Manifesto (Hicks et al., 2015). The goal of its ten principles, discussed individually below, is to reduce the chance that bibliometrics are used unwisely for research evaluation. The Metric Tide (Wilsdon et al., 2013) is more UK focused, and its 20 points cover wider system issues without adding research evaluation principles not included in the Leiden Manifesto. A 2022 review had a similar focus and also did not add to the principles, although it included the recommendation that all stakeholders should co-design research assessments (Curry et al., 2022). Finally, the CoARA Agreement on Reforming Research Assessment (coara.eu/agreement/the-agreement-full-text) from 2022 contains ten main points but only the first four cover the details of research assessments. Point 3 cautions against inappropriate indicator use, mentioning the h-index (covered implicitly by the Leiden Manifesto point 7) and Journal Impact Factors, which the Leiden Manifesto criticises but does not explicitly recommend against. CoARA point 4 cautions against using commercial university rankings in any assessment. Thus, overall, the Leiden Manifesto is the best starting point for a discussion of responsible use of indicators to support research evaluation. It is discussed in detail in the next section.

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; Zhang et al., 2017).

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 produced initiatives to combat them. With the rise of AI support for research evaluation, potential new problems must also be considered.

Leiden Manifesto principles: Relevance to LLM scores

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. This section discusses the relevance of each of the ten Leiden Manifesto principles for LLM scores, and any adjustments needed.

The ten Leiden Manifesto principles are arguably desirable goals for all research assessments. Nevertheless, there will be situations where they do not apply or cannot be achieved. For example, transparency may not be attainable if privacy is needed for an assessment, as in the UK REF2021.

Quantitative evaluation should support qualitative, expert assessment

Why? This principle is important because all quantitative indicators necessarily simplify the complex context of research, have limitations in accuracy, and add systematic biases towards whatever the indicator measures. For example, if citation-rates are used on their own then the system will reward high citation specialities, incentivising research into them. Using expert review safeguards against this as well as giving the potential to consider more information and context. Biases (in the form of a tendency for inappropriately higher or lower scores for one group, as defined above) are arguably more problematic than inaccuracy, given that human experts also differ, and so even expert review is imperfect at assessing academic research. Accuracy is more important at finer-grained levels of assessment.

LLMs. This principle is likely to be the same for LLMs as for other quantitative indicators, even though LLM scores mimic human peer review. This is important partly because LLMs do not evaluate scientific texts but only guess at their quality through advanced text pattern processing and inference. There may also be systematic biases in LLM scores.

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 favour males), 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 some fields get substantially higher average scores than others (Thelwall & Yaghi, 2025a), and some countries get higher scores, as do articles with longer abstracts (Thelwall & Kurt, 2025) but little else is known about any other disparities. Moreover, it is not clear whether these disparities are biases or reflect underlying quality differences.

Research 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 algorithms (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.

Measure performance against the research missions of the institution, group, or researcher

Why? This seems to be inherently desirable for all forms of research evaluation and so does not need an explicit justification. Nevertheless, it often seems to be ignored, perhaps with an implicit assumption that the goal is to evaluate research quality or excellence, or that whatever methods are adopted must be appropriate, especially if there are no practical alternatives. Some examples of missions or goals are briefly discussed here.

Goal: produce high quality research

Why? Rewarding high quality research seems to be the most common goal for research evaluations. The purpose might be to build scientific knowledge, attract and keep good researchers, or to increase reputation or improve rankings.

Expert review. If the evaluation goal is to assess research quality, then field experts would probably be widely recognised as the ultimate arbiters of this because fields form communities that manage their own standards and rules (Becher & Trowler, 2001). Thus, expert review is the optimal strategy. Nevertheless, experts disagree (e.g., Mališki & Mehmani, 2024), there may be insufficient specialist expertise to evaluate all the research and expert time is expensive. Indicators sometimes supplement experts because of the first two reasons (and to guard against human bias) and replace them because of the last one.

Citations. Citations are sometimes used to help evaluate research quality. Nevertheless, quality is usually thought of comprising originality, rigour and societal/scientific significance (Langfeldt et al., 2020) and citations primarily reflect scientific significance or visibility (Aksnes et al., 2019) so they are imperfect in this role. Since citation-based indicators correlate positively with research quality in most fields (Thelwall et al., 2023) they can still provide useful information. For example, they may help when experts disagree or for large-scale comparisons when it is reasonable to believe that their inaccuracies will average out enough that the results are useful, despite bias towards scientific impacts.

LLMs. LLMs can be configured to assess/guess research quality scores through their system instructions and so, in theory, they can cover the concept more completely than can citation-based indicators. This hypothesis is supported by evidence that their scores correlate more strongly with expert scores than do citation-based indicators in most fields (Thelwall, 2025a). If the goal is the generic one of assessing research quality, then the LLM prompt might ask the LLM to assess an article for the three core dimensions of rigour, significance and originality (Langfeldt et al., 2020). They would only be able to follow any instructions partially, however. For example, they may lack an ability to assess the rigour of mathematical proofs, conceptual frameworks, or statistical methods details.

Goal: produce societal impact or support local development

Why? Rewarding societal impact or local development is an increasingly common goal for research evaluations, such as with the Impact Case Studies component of the UK REF. This is a natural goal for governments seeking a national return or advantage from their investment in research.

Expert review. Expert and/or end user or stakeholder review logically seem to be the best choices for direct assessment of societal impact or local development. Nevertheless, assessing the societal impact of academic research may be difficult for non-academics, who may necessarily focus on their own industry sector or experience and who may not have expertise at evaluating academic research.

Citations. Citations would be a weak choice for these evaluations because they primarily reflect scientific influence.

LLMs. LLMs may have some capacity to assess societal impact or local development since LLM prompts could be explicitly tailored for assessing this. This has been shown to work reasonably well for UK Impact Case Studies in terms of correlating positively and moderately strongly with expert scores (Kousha & Thelwall, 2025).

Goal: produce scientific impact or value to science

Why? Rewarding scientific impact or value to science rather than research quality in general may be argued for on the basis that the point of science is progress and that there is little value in research that is not used. This seems to be a minority opinion for several reasons. First, if we only reward research that has scientific impact within the timeframe of an evaluation (e.g., reflected by citations) then we incentivise citation-chasing research, narrowing the scope of scientific activity. Second, publications also serve to certify the competence of researchers for appointments, grants, consultancy, reputation, and PhD supervision. Thus, rewarding more general research competence through publication is useful and can have longer term benefits in follow-up scientific impact. Third, it is difficult to predict which research areas will have future scientific value so it may be unhelpful to penalise people that guess wrong. For example, research into rapid vaccine development and airborne virus transmission became dramatically more important during 2020, but this importance could easily have emerged in 2010 or 2030 instead. Finally, and perhaps most importantly, the ultimate purpose of science is to help society so a focus on influence within science itself is ultimately pointless and risks encouraging inward-looking research (e.g., Tourish, 2020).

Expert review. If the evaluation goal is to reward contributions to scientific progress then review by field experts seem to be the authoritative choice because these organise fields (Becher & Trowler, 2001). Since experts disagree, using multiple experts and taking a consensus or average can increase accuracy. Of course, experts can have their own biases (e.g., toward their own field/approach/country/gender/friends), so any system using expert review would need to reduce the chance of this, such as through training or monitoring.

Citations. On the basis that citations can acknowledge prior work that the citing paper builds on (Merton, 1973), some argue that citations are better than peer review for recognising scientific impact or value to science (Rushforth & Hammarfelt, 2023). This is because each citation is potentially direct evidence of contributions whereas expert opinions are subjective and can be biased. Nevertheless, there are many reasons for citing, some of which are negative and many of which are relatively trivial (MacRoberts & MacRoberts, 2018; Seglen, 1998). Moreover, some work that is important to science does not get cited because it provides a definitive answer to a question so that there is no need for follow up. For example, suppose that a short weak paper introduces a question with an unknown answer (e.g., can LLMs provide useful research quality scores for journal articles?) or a wrong answer (e.g., cold fusion works). If the question is judged interesting, then the paper might get a high citation count from many follow-up papers but the article that answers the question definitively might get few or no citations because it has closed the topic. Similarly, methods contributions can get extremely highly cited because they become the standard choice for a popular technique even though there are similar alternatives (so the paper's unique contribution is minor), and some papers reporting demographic or other summary statistics (which are arguably not research in most fields) can be highly cited as background information for papers on a popular topic.

LLMs. LLMs probably have some capacity to assess scientific impact through prompts tailored for assessing it. Large-scale evidence of positive correlations between citation-based indicators and (overall) research quality scores (Thelwall, 2025a) suggests that this is likely, even though both are imperfect indicators of scientific impact.

A prompt designed purely for scientific impact would presumably perform better. This would be difficult to directly test in practice, however, since it would need a large-scale source of expert judgements about the scientific impact of articles separately from their rigour, originality, societal impact, and other properties. Currently the only large-scale evidence sources cover all dimensions simultaneously. Thus, whether LLMs can guess scientific impact better than citation-based indicators is likely to remain unknown for a long time.

Goal: increase scientific reputation

Why? Institutions and governments often seem to be more concerned with reputation than achievement in science, although the two are linked (Salter & Martin, 2001). A good national reputation for science can bolster confidence in a government, whereas a good university or country reputation for science can attract students, resources, and academic experts. In some cases, the link can be tangible and direct. For example, a position within the top 100 of a reputation-based university ranking can be necessary for some international mobility funding or for certain careers in some countries, even if the rankings lack validity. For this reason, some universities have university league table ranking positions as core mission targets and some departments and countries may consider this too.

Expert review. Reputation is inherently a subjective human property, so the opinions of the relevant target group are the gold standards unless a specific league table (however poor) is the mission target.

Citations. Citations may reflect reputation better than they reflect scientific impact based on the Matthew effect hypothesis that reputation attracts a disproportionate share of citations (Merton, 1968). Nevertheless, overall research quality may be more important for reputation than purely scientific impact (as primarily reflected by citations) and, as mentioned above, citations have limitations like rewarding early work triggering research more than later work definitively solving problems.

LLMs. As for scientific impact, LLMs probably have some ability to score research for reputation. Two advantages over citations are that they can be explicitly asked to assess reputation through customised prompts, and they seem to include huge amounts of text from the web in their training data (e.g., Almazrouei et al., 2023; Stryker, 2025). Thus, they may have insights into reputation outside academia, at least in theory.

Protect excellence in locally relevant research

Why? This seems to be desirable for all forms of research evaluation because locally relevant research is unlikely to be conducted far away. Thus, undervaluing locally relevant research will discourage it, reducing the value of academic research overall. For example, if analyses of international law are valued above analyses of national law, then a shortfall of research into national laws may result. In addition, locally relevant research may be necessary to translate or adapt global research findings for the local context to make it useful.

Expert review. Expert reviewers can be explicitly told to value locally relevant research despite a lack of international attention. In REF2021, for example, reviewers for broadly arts and humanities areas were cautioned that, "The terms 'world-leading', 'international' and 'national' will be taken as quality benchmarks within the generic definitions of the quality levels. They will relate to the actual, likely or deserved influence of the work, whether in the UK, a particular country or region outside the UK, or on international

audiences more broadly. There will be no assumption of any necessary international exposure in terms of publication or reception, or any necessary research content in terms of topic or approach." (REF2021, 2019, p. 47).

Citations**.** Citation-based indicators seem likely to be poor at this since they typically are international in scope, although there are exceptions like the Chinese Science Citation Database. Internationally relevant research therefore usually has a wider pool of potentially citing articles. This hypothesis does not seem to have been tested empirically, although international collaboration associates with more citations (Leydesdorff et al., 2019).

LLMs. LLM prompts can be explicitly tailored to request higher scores for locally relevant research, and there is some small scale evidence that this can work (Thelwall & Nunkoo, 2026).

Keep data collection and analytical processes open, transparent, and simple

Why? Open, transparent and simple processes allow those evaluated, or affected by an evaluation, to see and understand 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 addition, this can help with feedback to the assessed academics, helping them to improve and may be seen as desirable for natural justice (Colquitt et al., 2001; Heffernan & Flood, 2000). In practice, transparency is always partial. It may also not always be desirable, for example if public knowledge of low scores for individual researchers could cause them reputational or emotional harm (especially if the results are wrong) or might subject the evaluation system to large-scale legal attacks that would make it untenable. Thus, secrecy for parts of the process, such as the results, may be desirable. Of course, a deliberate lack of transparency is common in research: authors are rarely told the identities of the reviewers rejecting their papers or giving a low score to their grant applications (i.e., single/double blind peer review), and some decisions are made without any rationale.

Expert review. Expert review is arguably simple and can have elements of transparency if the names of the evaluators are known, the evaluation procedures are public, and the evaluators must submit a written justification of their scores, as in some forms of open peer review. In some but not all cases, these justifications may be reasonable and clear, such as by reporting that an article was given the lowest score because of a fundamental methodological mistake or a highest score because it solved a well-known problem. Nevertheless, many and possibly most research scoring decisions are probably subjective and not clear cut, with the associated justifications therefore being inherently imprecise. Moreover, no justifications could reveal the thought processes that led to their decisions and thought is in any case as complex as a human brain and at least partly subconscious and intuitive. For example, grant reviewers seem to be influenced by their emotional reactions to applications (Brunet & Müller, 2024), perhaps scoring those highest that they find most exciting.

Citations. The most transparent system seems to be citation-based indicators from the bibliometric database 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 for unknown reasons and with unknown influences. Other citation indexes have varying lesser degrees of transparency. Although the algorithms required to build and maintain a citation database are complex and cannot meet the simplicity criteria, citation-based indicator formulae can. The simplest citation-based indicator that works is preferable to a more complex one because those evaluated will be able to understand and check it more easily. In practice this rule favours

simple field normalised indicator formulae based on citation ratios (Waltman et al., 2011) rather than complex formulae or matrix-based iterative algorithms.

LLMs. LLMs have different transparency issues to peer review, and it is useful to compare them to get insights into both. Whilst the processes going on inside a reviewer's brain are invisible when they cogitate over what they have read and experienced when turning their knowledge into a score/judgement and report, the weights within an LLM that led to any score can be published if an open weights LLM (e.g., Hanke et al., 2024) is used. An Open Source LLM goes further by having transparent inputs (more so than a human reviewer): it is trained on a declared corpus. Thus, although the largest LLMs currently are not transparent, there are technologies that allow them to be. These seem to be minor advantages, however, given the overall complexity of even the smallest LLM, as discussed below.

The LLM input data collection process in terms of identifying the correct information for each article evaluated (e.g., its title and abstract) should be open, transparent, and simple as far as possible. This is not a trivial issue since there can be problems with different versions of articles, even those with DOIs. For transparency, the name and version of the LLM used can also be declared, as can the prompts and procedure to obtain the final grade (e.g., averaging five scores).

For analytical processes, all LLMs fail to be simple because of their high complexity. This complexity (e.g., at least 7 billion parameters for most small current LLMs) makes any model parameter information useless for anything except replicability. Thus, model transparency is not a practical advantage over peer review and is less helpful than for citations. Whilst it would be possible (and perhaps more practical/cheaper than with human reviewers) to use LLMs to generate reasoned score explanations to improve transparency it is not clear yet that these would be helpful. This because LLMs seem to be good at creating plausible explanations even for poor decisions (e.g., Du et al., 2024).

Allow those evaluated to verify data and analysis

Why? As for the above case, allowing the people evaluated to verify the data and analysis helps to build confidence in the evaluation system, to reduce errors, and may be fairer by protecting researchers to some extent from the adverse effects of errors.

Expert review. For expert review this might mean the chance to check that the correct research was evaluated (i.e., the data) and/or to post a rebuttal to an evaluation (i.e., the analysis). The latter is sometimes used in grant review processes (e.g., for UK Research and Innovation, UKRI). Rebuttals may not be practical for large-scale article-level evaluations, however.

Citations. For citations this seems to mainly entail allowing those evaluated to check for errors, such as missing citations, incorrectly applied formulae, or incorrectly classified articles.

LLMs. The largest current LLMs and some open weights LLMs arguably fail this for analysis processes in the sense that they do not publish their data sources (e.g., the documents used in their original training data), and these influence how scores are given. As mentioned above, published training data is a requirement for an LLM to be described as open source, rather than open weights (de Gregorio, 2025; Sowe et al., 2024). Data validation in the sense of the input data (e.g., article title and abstract) should be available. Moreover, to check for technical errors, it would be useful to allow those evaluated to replicate the actions of those obtaining the scores. For this, the evaluators should publish their

prompts, the exact identity of the LLM used (model name and version), any special parameter values and the random seed used. Although random seeds are not currently supported in the main web interfaces of the largest LLMs, they may be accessed through their APIs. Without specifying the random seed, identical inputs can lead to different outputs, preventing any analysis verification.

Account for variation by field in publication and citation practices

Why? Academic fields naturally vary substantially in publication and citation practices. Whilst researchers in some fields may routinely publish tens of descriptive articles annually for newly-identified life forms, others may typically require years to complete a single study, such as for a complex cell biology interaction. Similarly, whilst articles in hierarchical fast-moving subjects may rapidly accumulate high citation counts, papers in non-hierarchical deliberative and monograph-based fields may slowly reach low numbers. Not taking these differences into account would be unfair and push researchers into researching fields that have advantages.

Expert review. For expert review, this issue is normally dealt with as a natural side-effect of selecting evaluators from the field evaluated so that they understand the research norms. Differing publication speeds in multi-field evaluations can be at least partially dealt with by measures that do not advantage high publishing rates, such as by setting a maximum number of outputs to be evaluated, as in the UK REF.

Citations. For citations this means using field (and year) normalised indicators (Waltman et al., 2011) and avoiding them altogether for field where they are irrelevant.

LLMs. Users of LLMs should consider variations between fields in the average LLM scores (Thelwall, 2025a). This issue is not well understood yet, but it would be intuitively helpful to normalise scores when comparing between fields.

Base assessment of individual researchers on a qualitative judgement of their portfolio

Why? Although errors and biases (i.e., systematic errors favouring one group) in all quantitative research indicators might average out over large sets of documents, individual researchers have few, increasing the chance of substantial errors from indicators. Moreover, a researcher's contribution to science may be invisible in publications or citations. To give an extreme example, a female Chemistry Nobel Prize winner might devote the rest of her career to giving talks encouraging girls and women into science careers, a contribution lacking publications and citations. More routinely, evaluations of individual researchers need to consider their contributions to collaborative work as well as funding, service work, and other outputs, from datasets and software to public interest blogs.

Expert review. Expert review seems ideal to consider and balance all the contributions of an individual. The thoroughness and speed of a review may depend on the importance of the evaluation and the number of individuals to be assessed, however. For example, if there were 250 applicants for a position then a rapid triage may be necessary to remove unpromising applicants based on a quick scan of their CVs.

Citations. Citation-based indicators would be poor for individual researcher evaluation, as discussed above. Nevertheless, they might provide a source of supporting evidence of scientific impact.

LLMs. It is not clear whether LLMs have any capacity to make overall evaluations of an individual researcher's contributions, for example by scoring applicants from their CV or application letter. The most useful role might be at the triage stage for large number of applicants, perhaps as a second opinion after expert triage or as an initial triage decision maker, supported by a subsequent expert cross-check. Article-level LLM scores might also be harnessed to play a similar supportive role to citation-based indicators.

Avoid misplaced concreteness and false precision

Why? This is to avoid misleading those assessed by giving an unrealistic impression that the assessment method directly measures the quantity assessed (concreteness) and of the accuracy of the evaluation. Both apply to all forms of assessment, but the concreteness aspect varies between them.

Expert review. For concreteness, expert review evaluations should be transparent about the fact that experts disagree, make mistakes, and may not have the expertise to assess everything that they are required to.

Citations. Evaluations using citations should be clear about the relationship between the citation-based indicators and the evaluation goals. For all purposes this should include admitting that citations do not "measure" the phenomenon assessed but are only indicators in the sense that they may statistically associate positively with it. As mentioned above, Merton's (1973) theory posits that citations are scholarly acknowledgements of prior work that has aided the creation of new research, and this, if true, might justify a claim that citations measure research impact. The theory 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).

LLMs: LLM scores do not have the theorised direct (but partial) connection to research progress that citations do through Merton's (1973) theory. This perhaps makes it easier to avoid false claims of concreteness. It is still important to emphasise that the LLM scores, like citation rates, are only indicators and that they are used not because they are true but because they statistically associate with the theoretical goals being assessed (e.g., research quality). Trustworthiness is a common theme of AI research (Zhang et al., 2024), including being clear about the system capabilities and the purpose of the results. It is also important to acknowledge that LLMs can produce hallucination-like completely wrong answers, such as by giving a high score to an article that all expert reviewers would easily dismiss as wrong, fake or irrelevant. In addition, scores may vary greatly between LLMs and prompts, so it should be acknowledged that no score is the definitive LLM recommendation.

Recognize the systemic effects of assessment and indicators

Why? Whenever people are evaluated and know the evaluation method, some may 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 research evaluation approach might be ruled irresponsible if it generates perverse incentives.

Expert review. The guidelines for the reviewers might give the strongest steer as to the research properties that are most valued. These need to be designed so that researchers targeting them change their behaviours in ways that are positive for the system. For example, if the reviewer guidelines emphasise rigour, originality and significance then academics might strive for these. Perverse incentives might be generated if the reviewers are known to be biased and there may be no incentive if the reviewers are not trusted.

Citations. Using citation-based indicators would incentivise high citation specialties. Using field normalised citation-based indicators would avoid incentivising entire high citation fields but not specialties within fields. It would also incentivise research targeting scientific rather than societal benefits, and network-based gaming in the form of eliciting citations from others (e.g., citation cartels, editorial or refereeing self-citation encouragement).

LLMs. Similarly to expert review, the prompts used for LLM evaluations might provide the strongest steer to researchers about how to change their behaviour to improve their scores. LLMs may provide little incentive if they are not trusted, and known biases may also provide perverse incentives. In addition, authors may try to game the system by crafting 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 suggest rewrites to make it more likely to achieve a high score. This might involve exaggerating the importance of the findings to make the paper more like a press release (Thelwall, 2025b). This would waste time on an unproductive activity, but it is not clear that it would work. Journal peer review guards against unsupported claims, so LLM-friendly papers might be 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. Nevertheless, some authors may try (e.g., Sugiyama & Eguchi, 2025), and this should be considered when LLMs are used.

Scrutinize indicators regularly and update them

Why? The research system evolves over time, as does our understanding of good evaluation practices. Data sources may also change, and new ones emerge, so it is important to periodically revisit evaluation systems to ensure that best practice is still being followed.

Expert review. Reviewer updates might include change in personnel, procedures and guidelines.

Citations. Indicator formulae and citation indexes might be changed as improved versions emerge.

LLMs. This issue seems most relevant to LLMs, with new ones emerging regularly and updated models of existing LLMs being released multiple times per year. Thus, testing probably needs to be ongoing to ensure that the optimal LLM is selected. In addition, new ways of using LLMs may be developed, such as through different prompts or LLM-based agent systems, that would give improved results. For these purposes, it would be very useful (albeit challenging) to create benchmark datasets of articles with expert evaluation scores that LLMs could be tested against.

Principles not in the Leiden Manifesto

As discussed above, responsible uses of LLMs should consider the same issues as for bibliometrics. This section discusses some additional considerations for LLM-based scores. The first seems equally relevant for all, however.

Conduct a cost–benefit analysis of possible assessment methods

Why: All research assessment methods have costs in terms of human time for researchers, reviewers and/or managers if not for software and data. The theoretically best method may be impossible or irresponsible either because it is unaffordable or costs too much relative to the value of the assessment. Thus, it can be a pragmatic decision to avoid using the best methods due to their costs. A cost–benefit analysis should be conducted if there are reasonable grounds to believe that this is likely.

Expert review. Expert review is probably the most expensive method, whether in terms of payments or time. This is because considered expert review decisions are not quick. Assessments like the REF are costly (£471 million overall with £24 million spent on the expert review process) because of this (Technopolis, 2023). These costs are reasonable because of the amount of money directed by it (£15 billion) and its other benefits (system-wide accurate performance information). The cost of £7,000 per researcher evaluated makes this process prohibitively expensive for less important tasks. Moreover, countries with smaller research systems, because there are fewer inhabitants or they invest less in research, may need to pay international reviewers to fairly assess their work, which would be more expensive and probably less accurate due to weaker knowledge of local conditions. Whilst REF-type exercises are extreme in terms of scale, a similar logic applies to smaller-scale evaluations. Even at the level of individual academic appointments, some decision makers may lack access to the expertise or the resources to buy expertise necessary to evaluate candidates effectively.

Citations. Citation-based approaches need data, software and expertise. Citation data may cost if commercial sources like Scopus, the Web of Science or Dimensions is used, although OpenAlex is a free alternative that may sometimes be suitable. Much citation analysis software is free, except if built into the analytics toolkits of citation data providers but there is still a need for expertise to select and manage the software, data and indicator formulae and to interpret the results. If this expertise is commissioned externally, such as from teams associated with the data provider, then this may add to the overall costs.

LLMs: LLM-based approaches also need data, software and expertise. There is currently less LLM expertise for research evaluation than bibliometric expertise. 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. Again, externally commissioned expertise may be more expensive, and it is also currently (January 2026) not available.

Comply with copyright law

Why: All research evaluations need to comply with relevant laws, including copyright laws.

Expert review: Managers will need to ensure that evaluators are legally able to access the reports that they evaluate. This is usually attainable by authors securely sharing their work with evaluators, even if it is not open access.

Citations: Citation analyses using commercial data may need to ensure that their licence covers their application and that any results published do not breach copyright.

LLMs: The copyright situation for LLMs is more problematic than for citations and expert review. Before research works, or parts of research works, are entered into a LLM it needs to be clear that copyright is not breached. This is important because copyright conditions can exclude processing by AI, so publisher copyright statements need to be checked for exclusion clauses. For example, "Elsevier material may not be reproduced in combination with an artificial intelligence tool (including to train an algorithm, test, process, analyze, generate output and/or develop any form of artificial intelligence tool), or to create any derivative work and/or service (including resulting from the use of artificial intelligence tools)" (Elsevier, 2025b). If the system input is just titles and abstracts, then these may have a copyright exception in some nations and for some purposes.

Some countries have text and data mining copyright exemptions or fair use/dealing allowances, both of which allow LLM use in some or all contexts. Here, fair dealing refers to uses of copyright materials that are judged allowable because they do not unduly infringe the rights of the copyright holder to profit from the material (Saw, 2023). The UK has a text and data mining for academic research exception to copyright (Gov.uk, 2025b, section: "Text and data mining for non-commercial research") but this does not cover non-research applications or commercial development. It may cover research evaluation purposes, however, if they can be argued to be research exercises in themselves. Japan and Singapore extend this exemption to commercial AI (Stephens, 2024; Warren & Grasser, 2025) and in other countries, including the USA and Canada, it may be allowed by fair dealing legislation. Publisher permission seems to be currently (January 2026) needed in China, however.

If the system could learn from the inputs, then a second check should be made to see if this might breach copyright by the LLM recycling the inputs to train its model for future prompt responses for other users. The second issue can usually be avoided by using open source LLMs run locally or interfaces to LLMs that guarantee not to learn from inputs. For example, at the time of writing ChatGPT does not learn from anything submitted through its API. The UK has a copyright exemption for abstracts, "Where an article on a scientific or technical subject is published in a periodical accompanied by an abstract indicating the contents of the article, it is not an infringement of copyright in the abstract, or in the article, to copy the abstract or issue copies of it to the public." (Gov.uk, 2025a, Sect. 60) but this may not cover reuse for training by LLMs.

Most recently-published articles seem to be Open Access (OA), which influences the copyright situation. By 2023, according to Dimensions, 64% of articles were OA (COAlition S, 2025) and this would be substantially higher for countries like the UK with OA requirements or incentives. The common Creative Commons (CC) open access licences (by, by-nc, by-nc-nd, by-nc-sa, by-nd, by-sa: creativecommons.org/licenses/list.en) all permit non-commercial use, including by AI, although these all seem to prohibit use

within systems that learn from them because attribution could not be guaranteed. Thus, researchers, government and non-commercial research organisation seem to be able to safely use CC licenced research in non-learning LLM systems/interfaces, but this seems to prohibit use by for-profit research consultancies. CC licences seem to cover most open access research. Nevertheless, self-published green Open Access articles might be posted to the web without a licence instead of in an institutional or subject repository with built in licence support. These may still be copyright by default, but the purpose of copyright is primarily to prevent reproduction. The legal status of entering this copyright content into AI system to learn from it is currently contested but it seems safer when the content is entered to be scored rather than learned from since this involves no form of reproduction that leads to any form of republishing.

For non-OA articles, publisher copyright statements would need to be checked. Elsevier's subscription model copyright statement does not permit commercial text and data mining of subscription articles, but non-commercial non-sharing text and data mining is permitted, "Where legal access is obtained by a user, that user is able to text or data mine subscription articles for non-commercial purposes without sharing any adaptation of the original content with others" (Elsevier, 2025a). Publisher restrictions sometimes rule out training a system and this would prevent generating new systems through machine learning or fine-tuning existing systems but not processing the content with an existing system that does not learn from it. This would still allow processing by LLMs that were either private or did not learn from their inputs. As an alternative, authors may be able to submit preprints for processing, but often not the author accepted manuscript, which publishers may retain text and data mining options for.

A LLM research evaluation difference that does not need to be a principle

This section discusses one difference between LLMs and expert review and/or citation-based indicators that does not seem to be important enough to be added as a principle. It may influence the willingness of researchers and academics to accept an evaluation system.

The ephemerality of scores and differences between LLMs

Why not. Some evidence of the value of academic work is ephemeral. For example, an author judged to have 20 excellent papers by one LLM (or set of reviewers) might next year be judged to have only five by a different LLM or LLM version (or set of reviewers). It may be demoralising for scholars 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. Nevertheless, this mirrors the general academic situation in the sense that a person's work might go out of fashion, become obsolete or be ignored as attention focuses on other topics. Moreover, and more importantly, ephemerality does not seem to be a problem for research evaluation practice, since outputs are often seriously evaluated only once. For example, in the UK REF a different set of outputs gets scored in each iteration, so no output has scores that change between iterations. When outputs are evaluated multiple times, as might occur for job applicants with different universities evaluating their work, then it seems likely that Leiden Manifesto Principle 7 about individual researchers would override any ephemerality concerns. This has not been added

as a principle (although it was in the original paper) because it is primarily a consideration about whether researchers will accept an evaluation system than about how to design one responsibly.

Expert review. Expert scores are usually ephemeral in the sense that, with a few exceptions (F1000Prime, journals with public reviewer or editorial scores for articles), they are not publicly available and permanent but generated when needed for a particular assessment.

Citations. Ephemerality is a minor concern because citation counts are cumulative, decreasing only in response to citation database changes or indexing algorithm updates. Thus, scholars can track their "progress" or cumulative scientific recognition through their citation counts. Others can do the same, so that citation counts (e.g., for individual articles or via Google Scholar profiles) can be part of the evidence for scientific reputations.

LLMs. As argued above, LLM scores are ephemeral, but this would probably not be a concern for research evaluation. It might be a concern for researcher self-monitoring or reputation building through citations, if LLMs gain authority for research quality. This is more of an argument that citations are likely to persist as reputation evidence than that LLMs have limitations for individual research evaluations. As mentioned above, quantitative indicators should only be used to support qualitative judgements for individual researcher evaluation (Hicks et al., 2015).

Summary

Research evaluation processes now have a choice between expert review, bibliometrics, and/or AI/LLMs. The decision about the approach to follow should be guided by accuracy, fairness and desirable systemic outcomes. The Leiden Manifesto principles provide a recognised framework for this, although two extensions are needed, both of which have become more important in the context of LLMs (Table 1). This article has argued for the importance of all twelve, considering how each one might be relevant for expert review, citation analysis and LLM scores. Of course, every evaluation has special conditions and may not be able to meet all principles due to valid exceptions. For example, a cost–benefit analysis (number 11) may indicate that expert evaluation (number 1) is unaffordable in a context where the decisions to be made are not important enough to justify the time of the experts. The extended Leiden Manifesto principles are nevertheless intended to guide future research evaluations, and especially considerations about which approach to use when LLMs are a possible choice.

Returning to the LLM focus of this article, it seems clear that all the main responsible research evaluation considerations for bibliometrics also apply to LLMs but need to be adapted. For example, transparency in the context of LLMs might entail declaring LLM names and versions as well as the prompts and the score processing algorithms. Even with this, transparency causes a particular concern for LLM evaluations because the decision-making processes are too complex to understand.

The need to consider the financial or time cost of the different potential evaluation methods is a new practical/responsible consideration. It applies to all evaluation methods and is perhaps most important in low resource contexts. It should also be considered when criticising evaluation systems that violate some of the principles. Finally, as illustrated by the discussion of the twelfth principle, in some countries the most important practical/responsible limitation of LLMs for research evaluation may be the need to comply with

Table 1 Leiden Manifesto rules (1–10) and LLM-motivated extensions (11–12)

No	Principle (see Sects. "Quantitative evaluation should support qualitative, expert assessment" to "Scrutinize indicators regularly and update them", "Conduct a cost–benefit analysis of possible assessment methods" and "Comply with copyright law" for explanations)	Applies to*
1	Quantitative evaluation should support qualitative, expert assessment	Cites, LLMs
2	Measure performance against the research missions of the institution, group, or researcher	All
3	Protect excellence in locally relevant research	All
4	Keep data collection and analytical processes open, transparent, and simple	Cites, LLMs
5	Allow those evaluated to verify data and analysis	All
6	Account for variation by field in publication and citation practices	All
7	Base assessment of individual researchers on a qualitative judgement of their portfolio	Experts
8	Avoid misplaced concreteness and false precision	Cites, LLMs
9	Recognize the systemic effects of assessment and indicators	All
10	Scrutinize indicators regularly and update them	Cites, LLMs
11	Conduct a cost–benefit analysis of the possible assessment methods	All
12	Comply with copyright law	All (esp. LLMs)

*Cites = citation-based indicators; LLMs = LLM-based scores; Experts = Expert review scores; All = all forms of research evaluation, including expert review

local copyright law, given that a large collection of documents may take considerable time to check or obtain compliance for if there is no legislation that covers them all.

Prior publication

This paper is based on, and has some text reused from, "Responsible Uses of Large Language Models" published in Proceedings of the 20th International Society of Scientometrics and Informetrics Conference (ISSI2025) Volume 1. (pp. 71–80).

Funding The first author is a member of the Distinguished Reviewers Board of this journal and was funded by an ESRC Metascience (UK) grant (APP43146).

References

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

Aksnes, D. W., Langfeldt, L., & Wouters, P. (2019). Citations, citation indicators, and research quality: An overview of basic concepts and theories. *SAGE Open, 9*(1), Article 2158244019829575.

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.

Almazrouei, E., Alobeidli, H., Alshamsi, A., Cappelli, A., Cojocaru, R., Debbah, M., & Penedo, G. (2023). The falcon series of open language models. arXiv preprint arXiv:2311.16867

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

Becher, T., & Trowler, P. (2001). *Academic tribes and territories*. McGraw-Hill Education.

Brunet, L., & Müller, R. (2024). The feeling rules of peer review: Defining, displaying, and managing emotions in evaluation for research funding. *Minerva, 62*(2), 167–192.

COAlition S (2025). Annual Review 2024: Accelerating Open Access. https://www.coalition-s.org/wp-content/uploads/2025/04/Plan-S-2024-annual-review.pdf

Colquitt, J. A., Conlon, D. E., Wesson, M. J., Porter, C. O., & Ng, K. Y. (2001). Justice at the millennium: A meta-analytic review of 25 years of organizational justice research. *Journal of Applied Psychology, 86*(3), 425.

Curry, S., Gadd, E., & Wilsdon, J. (2022). *Harnessing the Metric Tide: Indicators, infrastructures & priorities for UK responsible research assessment*. Loughborough University.

Dahler-Larsen, P. (2011). *The evaluation society*. Stanford University Press.

Du, J., Wang, Y., Zhao, W., Deng, Z., Liu, S., Lou, R., & Yin, W. (2024, November). LLMs assist NLP researchers: Critique paper (meta-) reviewing. In *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp. 5081–5099).

EC (2024). Science, Research and Innovation performance of the EU 2024 report. https://research-and-innovation.ec.europa.eu/knowledge-publications-tools-and-data/publications/all-publications/science-research-and-innovation-performance-eu-2024-report_en

Elsevier (2025a). End user reuse for content published by Elsevier. https://www.elsevier.com/en-gb/about/policies-and-standards/open-access-licenses

Elsevier (2025b). Permission request. https://www.elsevier.com/en-gb/about/policies-and-standards/copyright/permissions/permission-request-form

Gadd, E. (2020). University rankings need a rethink. *Nature, 587*(7835), 523–524.
Gov.uk (2025a). Copyright, Designs and Patents Act 1988. https://www.legislation.gov.uk/ukpga/1988/48/section/60
Gov.uk (2025b). Exceptions to copyright. https://www.gov.uk/guidance/exceptions-to-copyright
de Gregorio, A. (2025). Mitigating cyber risk in the age of open-weight LLMs: Policy gaps and technical realities. arXiv preprint arXiv:2505.17109.
Hanke, V., Blanchard, T., Boenisch, F., Olatunji, I. E., Backes, M., & Dziedzic, A. (2024). Open LLMs are necessary for private adaptations and outperform their closed alternatives. In ICML 2024 Next Generation of AI Safety Workshop.
Hcéres (2025). Department of research evaluation https://www.hceres.fr/en/department-research-evaluation
Heffernan, M. M., & Flood, P. C. (2000). An exploration of the relationships between the adoption of managerial competencies, organisational characteristics, human resource sophistication and performance in Irish organisations. *Journal of European Industrial Training, 24*(2/3/4), 128–136.
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*, Article 101861.
Kousha, K., & Thelwall, M. (2025). Assessing the societal influence of academic research with ChatGPT: Impact case study evaluations. *Journal of the Association for Information Science and Technology, 76*(10), 1357–1373. https://doi.org/10.1002/asi.25021
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.
Leydesdorff, L., Bornmann, L., & Wagner, C. S. (2019). The relative influences of government funding and international collaboration on citation impact. *Journal of the Association for Information Science and Technology, 70*(2), 198–201.
Liang, W., Izzo, Z., Zhang, Y., Lepp, H., Cao, H., Zhao, X., & Zou, J. Y. (2024). Monitoring ai-modified content at scale: A case study on the impact of chatgpt on ai conference peer reviews. *arXiv preprint* arXiv:2403.07183.
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.
Malički, M., & Mehmani, B. (2024). Structured peer review: Pilot results from 23 Elsevier journals. *PeerJ, 12*, Article e17514.
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. (1968). The Matthew effect in science: The reward and communication systems of science are considered. *Science, 159*(3810), 56–63.
Merton, R. K. (1973). *The sociology of science: Theoretical and empirical investigations*. University of Chicago Press.
Moed, H. F. (2005). *Citation Analysis in Research Evaluation*. Springer.
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.
REF2021 (2019). Panel criteria and working methods (2019/02). https://2021.ref.ac.uk/publications-and-reports/panel-criteria-and-working-methods-201902/index.html
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*, Article 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), Article 102946. https://doi.org/10.1016/j.dsx.2024.102946

Salter, A. J., & Martin, B. R. (2001). The economic benefits of publicly funded basic research: A critical review. *Research Policy, 30*(3), 509–532.

Saw, C. L. (2023). Distinguishing the fair use and fair dealing doctrines in copyright law—much ado about nothing? *Journal of Intellectual Property Law and Practice, 18*(12), 848–866.

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.

Sowe, S., Mou, Y., Cheng, D., Kong, L., Neumann, A. T., & Decker, S. (2024). Understanding Open Source Large Language Models: An Exploratory Study. *2024 2nd International Conference on Foundation and Large Language Models (FLLM)* (pp. 132–140). IEEE Press.

Stephens, H. (2024). Singapore's new copyright act three years on: There's no need to open the AI exception door even wider. https://hughstephensblog.net/2024/09/02/singapores-new-copyright-act-three-years-on-theres-no-need-to-open-the-ai-exception-door-even-wider/

Stryker, C. (2025). What are large language models (LLMs)? https://www.ibm.com/think/topics/large-language-models

Sugiyama, S. & Eguchi, R. (2025). 'Positive review only': Researchers hide AI prompts in papers. https://asia.nikkei.com/Business/Technology/Artificial-intelligence/Positive-review-only-Researchers-hide-AI-prompts-in-papers

Technopolis (2023). REF 2021 Cost Evaluation. https://repository.jisc.ac.uk/9184/1/REF_2021_cost_evaluation_final_report.pdf

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

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

Thelwall, M. & Nunkoo, R. (2026). A Global South strategy for evaluating research value with ChatGPT. Quantitative Science Studies.

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. (2025b). Research quality evaluation by AI in the era of Large Language Models: Advantages, disadvantages, and systemic effects. *Scientometrics, 130*(10), 5309–5321. https://doi.org/10.1007/s11192-025-05361-8

Thelwall, M., Jiang, X., & Bath, P. (2025). Estimating the quality of published medical research with ChatGPT. *Information Processing & Management, 62*(4), Article 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., Kousha, K., Stuart, E., Makita, M., Abdoli, M., Wilson, P., & Levitt, J. (2023). In which fields are citations indicators of research quality? *Journal of the Association for Information Science and Technology, 74*(8), 941–953. https://doi.org/10.1002/asi.24767

Thelwall, M., & Kurt, Z. (2025). Research evaluation with ChatGPT: Is it age, country, length, or field biased? *Scientometrics, 130*(10), 5323–5343. https://doi.org/10.1007/s11192-025-05393-0

Thelwall, M., & Yaghi, A. (2025a). In which fields can ChatGPT detect journal article quality? An evaluation of REF2021 results. *Trends in Information Management, 13*(1), 1–29. https://doi.org/10.48550/arXiv.2409.16695

Tourish, D. (2020). The triumph of nonsense in management studies. *Academy of Management Learning & Education, 19*(1), 99–109.

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), Article e0193762.

Waltman, L., van Eck, N. J., van Leeuwen, T. N., Visser, M. S., & van Raan, A. F. (2011). Towards a new crown indicator: Some theoretical considerations. *Journal of Informetrics, 5*(1), 37–47.

Wang, J., Shi, E., Yu, S., Wu, Z., Hu, H., Ma, C., & Zhang, S. (2025). Prompt engineering for healthcare: Methodologies and applications. *Meta-Radiology*, 100190.

Warren, S. & Grasser, J. (2025). Japan's New Draft Guidelines on AI and Copyright: Is It Really OK to Train AI Using Pirated Materials? https://www.privacyworld.blog/2024/03/japans-new-draft-guidelines-on-ai-and-copyright-is-it-really-ok-to-train-ai-using-pirated-materials/

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

Wu, M. J., Zhang, Y., Haunschild, R., & Bornmann, L. (2025, June). Leveraging large language models for post-publication peer review: Potential and limitations. In *Proceedings of the 20th International Conference on Scientometrics & Informetrics (ISSI 2025)* (pp. 1207-1226).

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.

Zhang, L., Rousseau, R., & Sivertsen, G. (2017). Science deserves to be judged by its contents, not by its wrapping: Revisiting Seglen's work on journal impact and research evaluation. *PLoS ONE, 12*(3), Article e0174205.

Zhang, Y., Wu, M., Zhang, G., & Lu, J. (2024). Responsible AI: Portraits with Intelligent Bibliometrics. *IEEE Transactions on Artificial Intelligence*. https://doi.org/10.1109/TAI.2024.3510474

Zhou, R., Chen, L. and Yu, K., 2024, May. 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).