Can AI assess research?

NIFU 13th August 2026, 13.30 -15.00

Agenda

1. **Welcome** by Fredrik Piro, Head of Research, NIFU
2. **How can LLMs best be used for the post-publication research quality assessment of journal articles?**
   Mike Thelwall, University of Sheffield
3. **Differences between human and LLM review reports**
   Dag W. Aksnes, NIFU
4. **Can LLMs be guided to review like humans?**
   Henrik Karlstrøm, NIFU
5. **Policy brief 'How (not) to use Large Language Models to assess research'**
   Liv Langfeldt, NIFU
6. **Open discussion/Q&A**

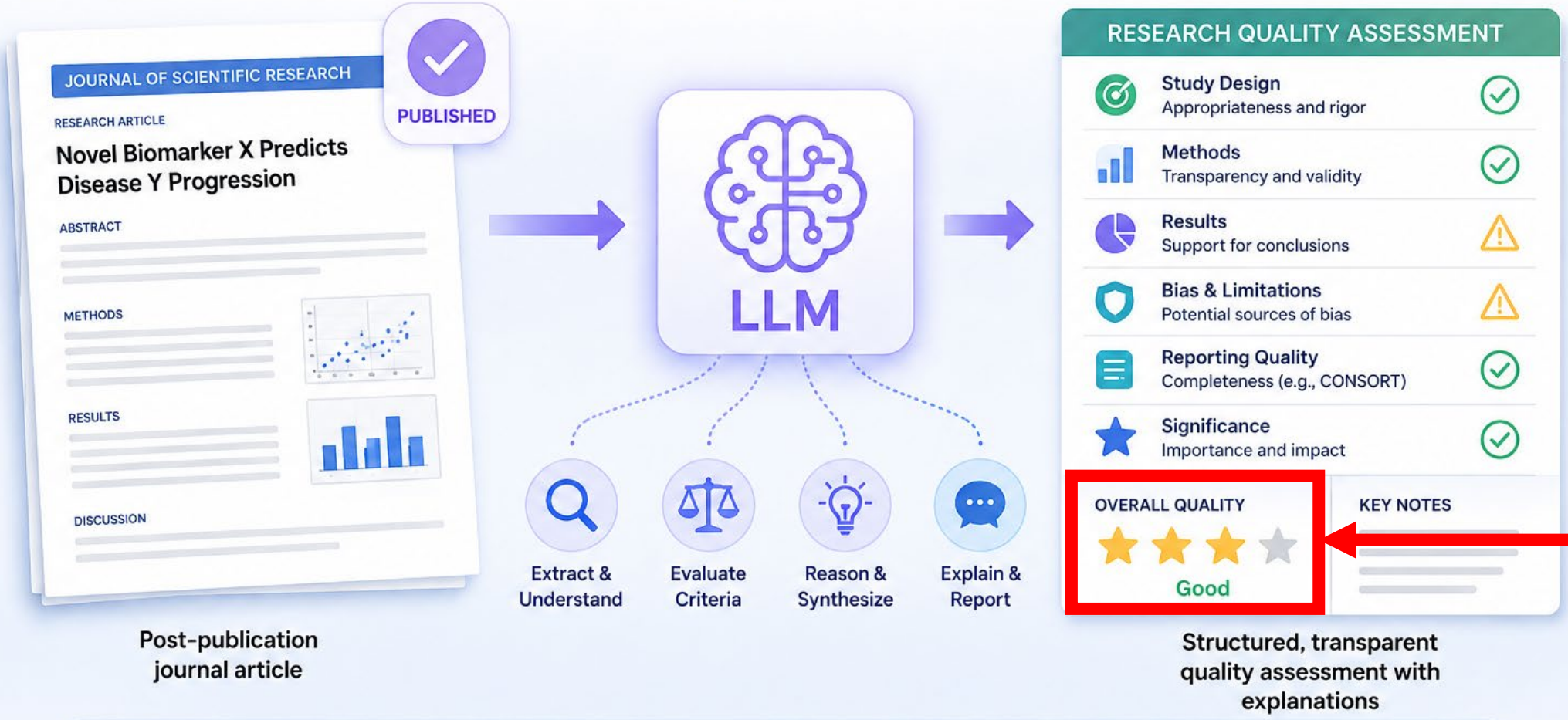
The AI Peer project
https://sites.google.com/sheffield.ac.uk/peer-ai/home

Policy Brief
NIFU Innsikt 2026:6
https://hdl.handle.net/11250/5555680

NIFU Nordic Institute for Studies in Innovation, Research and Education

Mike Thelwall, University of Sheffield

How can LLMs best be used for the post-publication research quality assessment of journal articles?



This talk

Post-publication journal article

Structured, transparent quality assessment with explanations

Scalable and efficient

Consistent and objective

Supports editors, reviewers, and readers

Improves research integrity and trust

ChatGPT prompt from the UK Research Excellence Framework (REF) 2021

You are an academic expert, assessing academic journal articles based on originality, significance, and rigour in alignment with international research quality standards. […]

Originality will be understood as the extent to which the output makes an important and innovative contribution to understanding and knowledge in the field. […]

Significance will be understood as the extent to which the work has influenced, or has the capacity to influence, knowledge and scholarly thought, or the development and understanding of policy and/or practice.

Rigour will be understood as the extent to which the work demonstrates intellectual coherence and integrity, and adopts robust and appropriate concepts, analyses, sources, theories and/or methodologies.

The scoring system used is 1*, 2*, 3* or 4*, which are defined as follows.

4*: Quality that is **world-leading** in terms of originality, significance and rigour.

3*: Quality that is **internationally excellent** in terms of originality, significance and rigour but which falls short of the highest standards of excellence.

2*: Quality that is **recognised internationally** in terms of originality, significance and rigour.

1* Quality that is **recognised nationally** in terms of originality, significance and rigour.

[…]

[]Based on the REF criteria for assessing research outputs, the article would be evaluated under the following categories:

1. **Originality**: The study demonstrates originality through its comprehensive, large-scale analysis of academic collaboration patterns across multiple countries and disciplines. It engages with the complex problem of understanding the dynamics of academic collaboration and its impact on citation rates, a topic of significant relevance in the context of global research practices.
2. **Significance**: The research has a significant potential to influence understanding in multiple fields, particularly in research policy and scholarly communication. It provides valuable insights that can inform national and international strategies for research collaboration and funding. The study's findings challenge the uniformity of collaboration practices across different regions and disciplines, highlighting the need for context-specific approaches in research management and policy-making.
3. **Rigour**: The study exhibits intellectual coherence and integrity. It adopts a robust and appropriate methodology, utilizing a large dataset and a well-justified metric for citation impact analysis. The research design is methodologically sound, and the authors provide a thorough analysis of their findings, acknowledging the limitations of their study.

Based on these considerations, the article would be rated as follows:

- **4* (**World-leading**)**: While the study is methodologically sound and offers new insights into the patterns of academic collaboration and citation impact, it may not reach the highest standards of world-leading work in terms of providing transformative theoretical advances or developing entirely new methods or paradigms. It offers substantial contributions to the field but may fall slightly short of a pioneering breakthrough that would characterize a definitive 4* rating.
- **3* (**Internationally excellent**)**: The article most likely fits within this category. It is of high quality, contributes important knowledge, and is likely to be influential at an international level. The study is well-designed, applies a robust methodology, and addresses a significant issue in academic research. It is likely to be an important point of reference for future research in this area.

In conclusion, the article is assessed to be of **3*** quality, indicating that it is internationally excellent in terms of originality, significance, and rigour.

ChatGPT writes plausible evaluations

ChatGPT

Experimental dataset 1

- I selected 51 of my own journal articles and gave each one a private quality score between 1* and 4*, including fractions.
- I gave the scores before testing with ChatGPT.
- I then compared ChatGPT's scores against mine.
- My scores have never been published so ChatGPT cannot cheat.

- Example query (not one of the 51 papers, not using the API, but showing the report format)

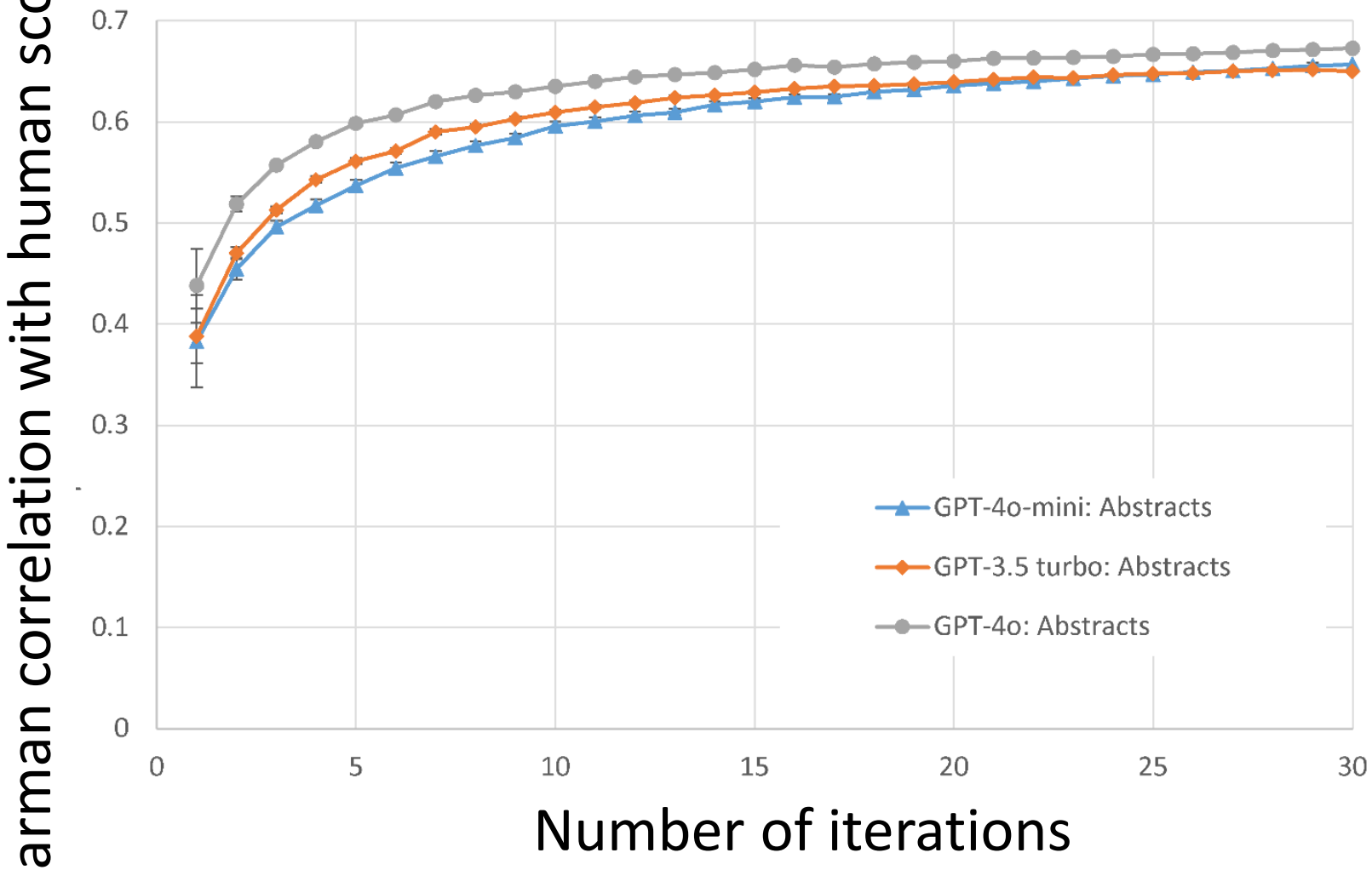
Initial findings

- ChatGPT's scores were usually wrong.
- It usually predicted 3* and rarely predicted 4* or 2* and almost never predicted 1*.
- The reports were *plausible* but the scores very *unreliable*.
- But they are still useful! ….

Average ChatGPT Scores against my scores for 51 of my articles: Input title and abstract

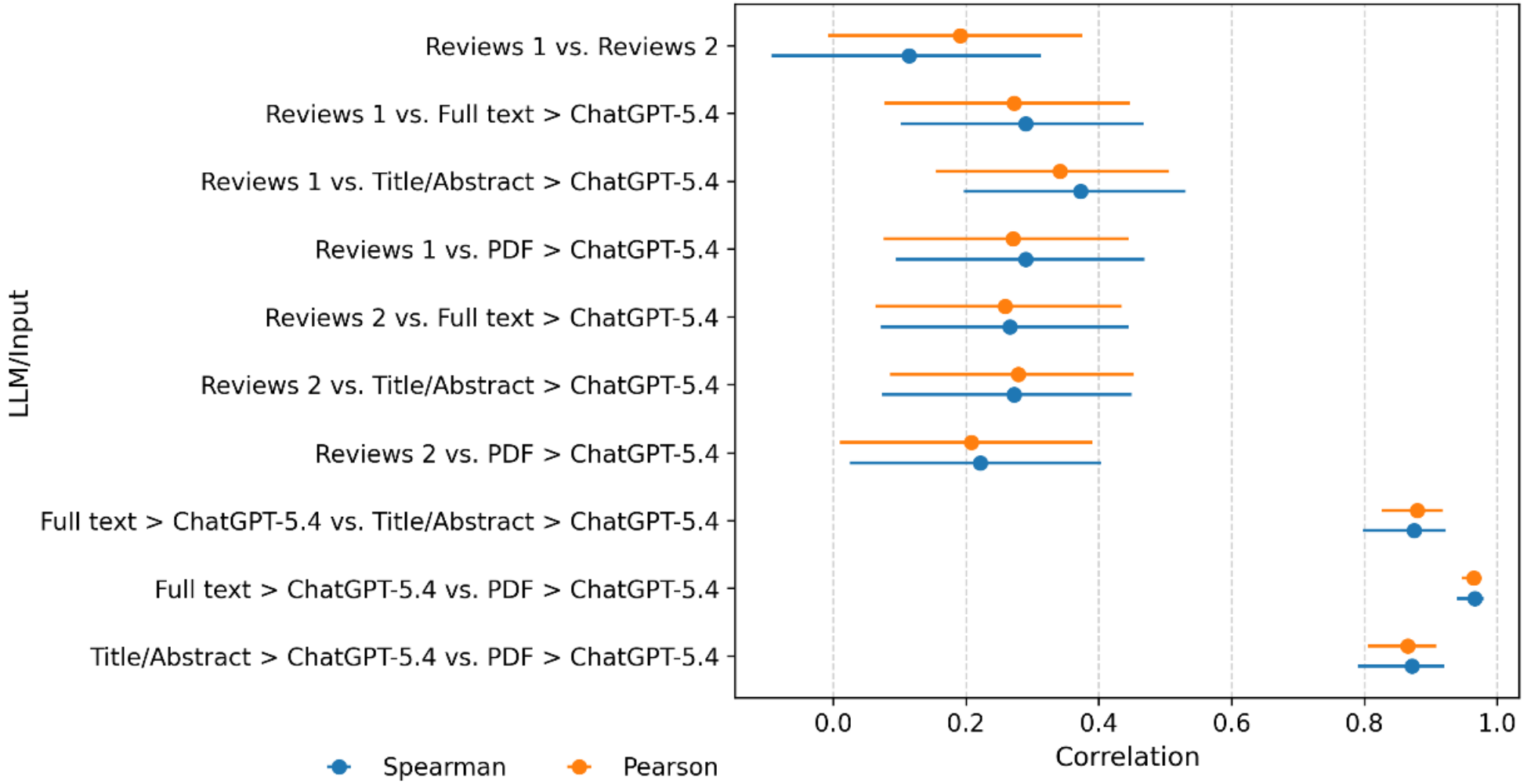
Strong Spearman correlation between my quality scores and ChatGPT averages for 51 of my papers.

Correlation *increases substantially* when scores are averaged.



My estimated REF score for 51 of my papers

Dataset 2: 98 journal articles within UoA 3 (Allied Health Professions, Dentistry, Nursing and Pharmacy). With University of Sheffield internal departmental review scores.



- Evidence that ChatGPT-5.4 gives worse results with article PDFs than if only shown titles and abstracts.
- For titles/abstracts, it agrees with individual departmental reviewers **more than the reviewers agree with each other**.

ChatGPT 4o

“Do squirrel surgeons generate more citation impact?”

“[…] The research utilizes a robust methodology, analyzing the field-normalized citation impact of 94,001 US surgery articles through linear regression, **accounting for various variables like author species, publication year, and author numbers**.

In terms of originality, **the work is highly innovative, engaging with a novel approach in citation analysis and addressing a unique aspect of species differences in academic research**. […] Given these considerations, the study could be rated as ****3* (internationally excellent)**.** […]” ChatGPT 4o

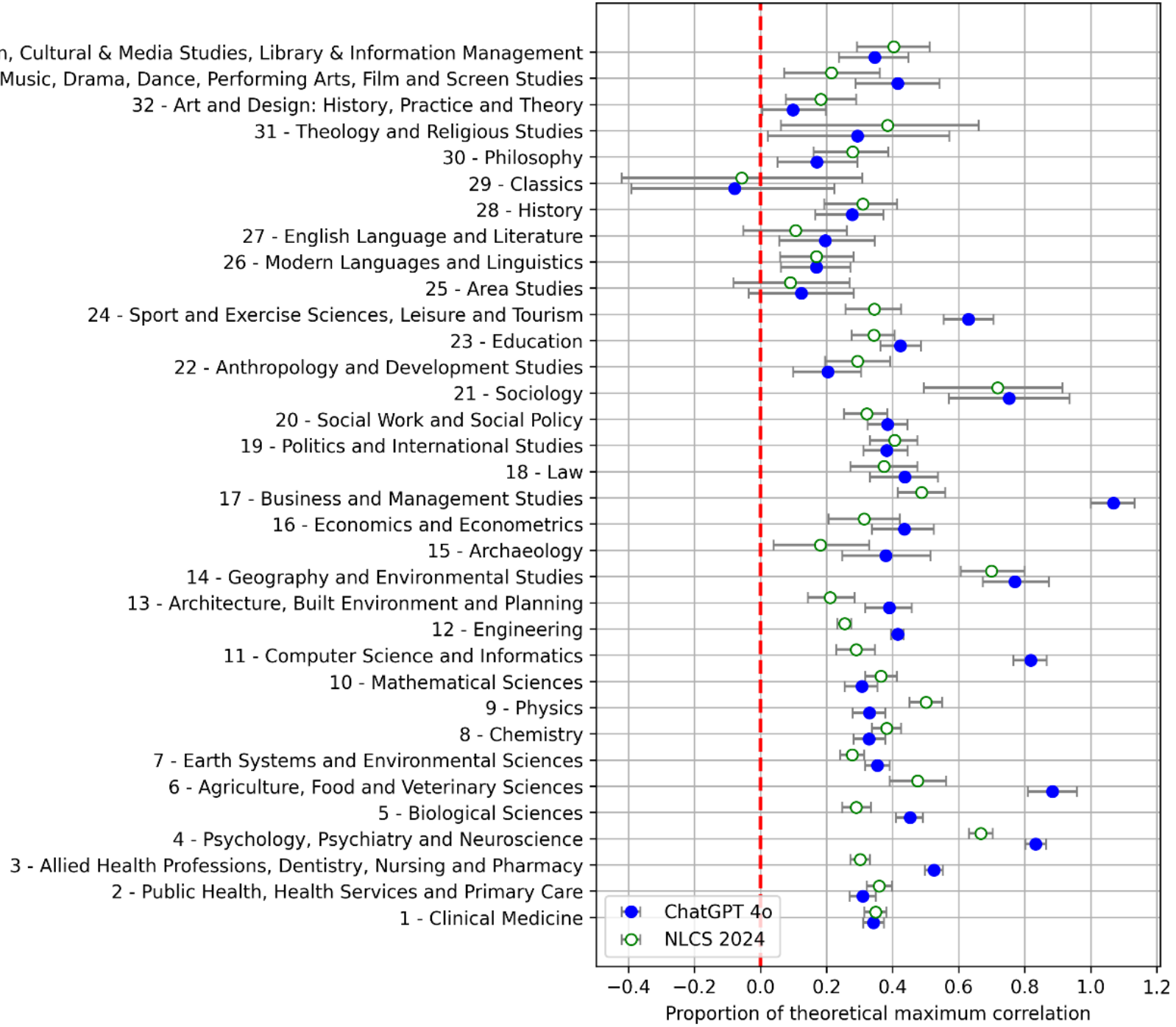
(sometimes 4*)

Experimental dataset 3 for larger scale tests

- I sampled 107,212 articles from the “Unit of Assessment” disciplinary groupings in the UK’s national Research Excellence Framework (REF) 2021.
- The REF2021 quality scores for individual articles are missing, so I replaced them with the average quality of the department that they are from.
- This is an indirect quality score.
- A positive correlation would suggest some ability of ChatGPT to assess research quality.
- The departmental average scores are “sort of” public and the connection between articles and departments is also “sort of” public so ChatGPT could cheat (but it does not seem to).

ChatGPT 4o vs. **mature** citations

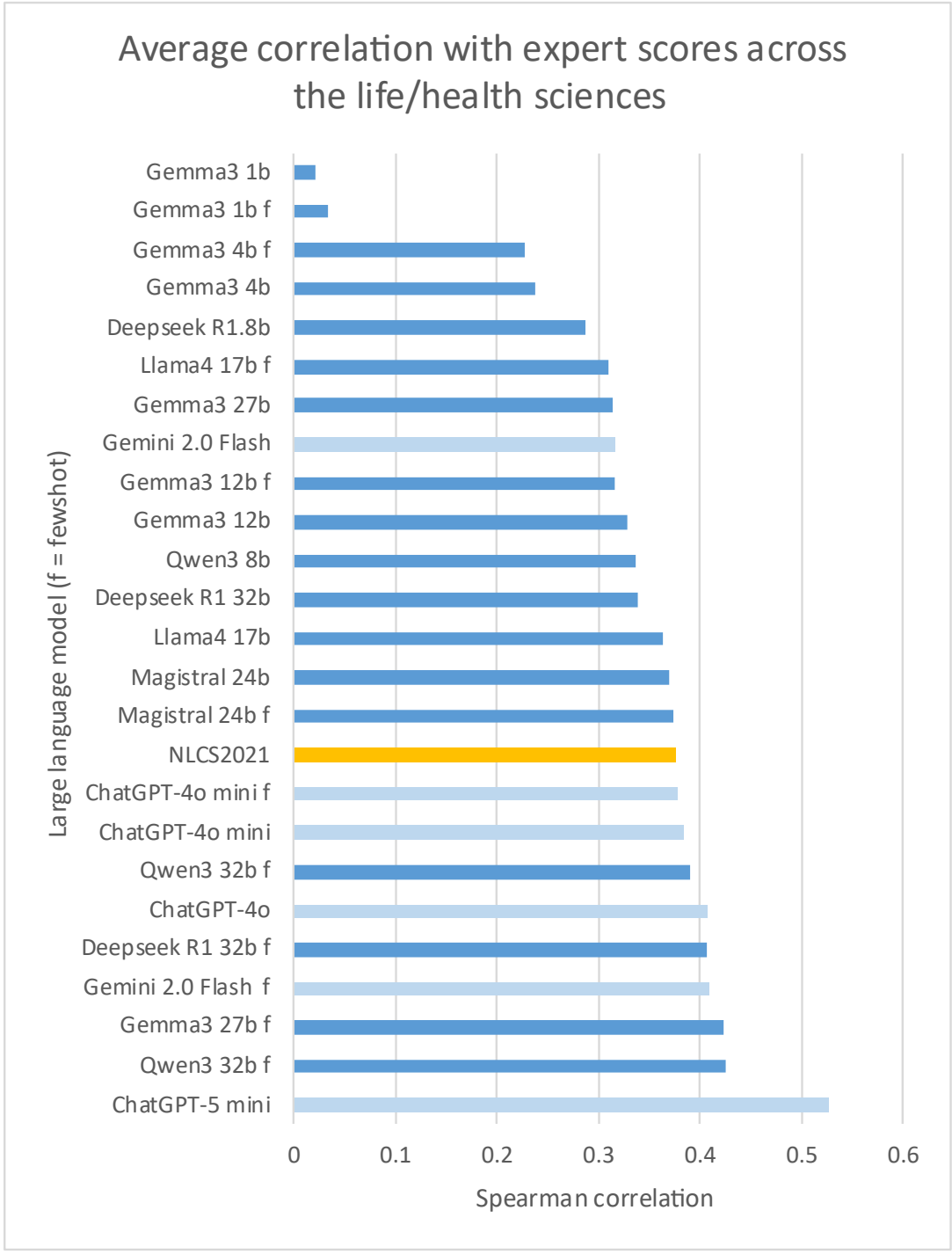
- ChatGPT 4o has a stronger correlation than citation rates (NLCS) for **medium term** citations with research quality in most fields.
- For this graph, the proportion of theoretical maximum correlation is shown.
- NLCS = Normalised Log-transformed Citation Score



Comparison between LLMs

All except the smallest (1b) LLMs seem to have some ability to guess research quality, but ChatGPT-5 seems to be the best.

All just based on the title and abstract!



Might ChatGPT be cheating by looking up article scores for dataset 3? I don't think so

- I fed article titles and abstracts to ChatGPT-4o mini API and asked it for the publishing journal, first author affiliation and year, but it was usually wrong, even for highly cited articles.
- So it does not have a strong ability to recognize articles from titles and abstracts.

Article set	Articles	Scopus - ChatGPT matches		
		Journal	Affiliation	Year
Uncited	951	2.7%	6.7%	29.8%
Energies	7976	0.3%	4.5%	23.4%
Physical Review B	5048	41.7%	1.8%	20.1%
Frontiers in Psychology	5408	4.0%	4.3%	29.3%
PLoS One	15034	3.5%	9.7%	31.2%
IEEE Access	11593	2.9%	2.5%	24.4%
ACS Applied Materials & Int.	5832	2.5%	2.0%	17.6%
Chemical Engineering Journal	4456	6.2%	1.9%	11.9%
Nature Communications	6778	18.4%	4.1%	26.1%
Highly cited	979	32.6%	13.6%	48.4%

Thelwall, M. (2026). Do Large Language Models know basic facts about journal articles? *Journal of Documentation*. https://doi.org/10.1108/JD-11-2025-0330

Summary

- LLMs and particularly ChatGPT seem to have a greater ability to guess research quality than citation-based indicators – run 5+ times *and rank (or scale) the results* to get more reliable information.
- Only use to **rank** articles for large scale evaluations not to score individual articles.
- The accuracy of the rankings can match individual human accuracy in some cases.
  - See Liv's and Dag's talks for recommendations

References and bibliography

1. Thelwall, M., & Ali, P. (2026). Is ChatGPT as reliable as individual reviewers assessing the quality of published journal articles from PDFs or titles and abstracts?. *arXiv preprint arXiv:2607.25965*.
2. Langfeldt, L., Aksnes, D. W., Karlstrøm, H., & Thelwall, M. (2026). Large Language Models for departmental expert review quality scores. arXiv preprint arXiv:2601.18945.
3. Thelwall, M. (2026). In which fields do ChatGPT scores align more closely with research quality than do citation rates? *Journal of Data and Information Science*. https://doi.org/10.1515/jdis-2026-0058
4. 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
5. Thelwall, M. (2024). *Quantitative Methods in Research Evaluation: Citation Indicators, Altmetrics, and Artificial Intelligence*. https://arxiv.org/abs/2407.00135 (book preprint)
6. Thelwall, M. (2025). Can smaller large language models evaluate research quality? *Malaysian Journal of Library and Information Science*, 30(2), 66-81. https://doi.org/10.22452/mjlis.vol30no2.4
7. 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
8. Thelwall, M. (2025). 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
9. Thelwall, M., &, 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
10. Thelwall, M. (2025). Evaluating research quality with Large Language Models: An analysis of ChatGPT's effectiveness with different settings and inputs. *Journal of Data and Information Science*, 10(1), 7-25. https://doi.org/10.2478/jdis-2025-0011
11. Thelwall, M. (2026). Prompt perturbation and fraction facilitation sometimes strengthen Large Language Model scores. Data and Information Management, 10(3), 100140. https://doi.org/10.1016/j.dim.2026.100140
12. Thelwall, M. (2026). Large Language Models and responsible research evaluation: An extension of the Leiden Manifesto. *Scientometrics.* https://doi.org/10.1007/s11192-026-05552-x

Differences between human and LLM review reports

Dag W. Aksnes

NIFU seminar 13. August 2026

NIFU

Nordic Institute for Studies in Innovation, Research and Education

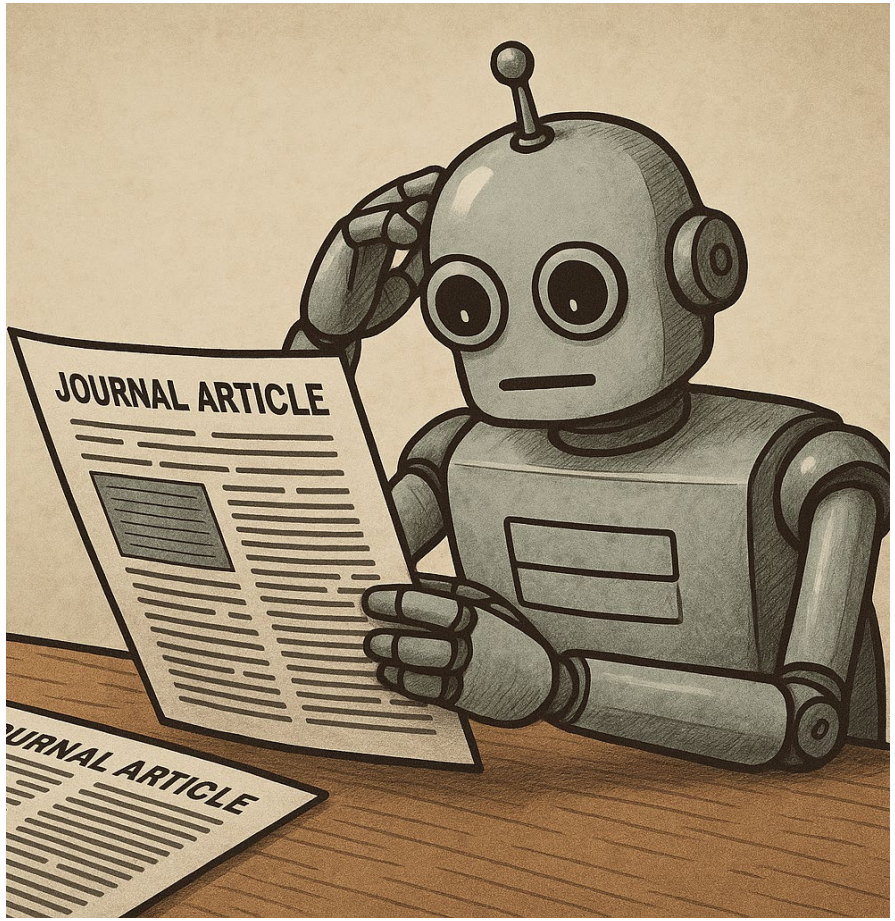
Introduction

Based on selected findings from the AI peer project.

How LLM scores correspond with human scores

What kinds of evaluative signals LLMs use when generating journal article reviews

Are they reading like a domain expert, or are they doing something quite different?

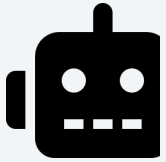


APPROACH

Comparing article reviews

Journal articles

Sample: Journal articles selected for departmental reviews



Submitted to LLMs

Full text and title/abstract inputs

Human expert reviews

Same articles, departmental review

Compared

Scores and assessments side by side

Note on sample: Prompt = REF2021 evaluator instructions (adapted). Articles are not a random sample — they were preselected as their best outputs, in line with REF-style submission practice.

The REF scoring framework (1*–4*)

4*

World-leading

Quality that is world-leading

3*

Internationally excellent

Quality that is internationally excellent but which falls short of the highest standards of excellence

2*

Recognised internationally

Quality that is recognised internationally

1*

Recognised nationally

Quality that is recognised nationally

Assessed against three criteria: originality, significance, and rigour — each scored using the same generic definitions applied to every article.

CONTEXT

A moving target: LLMs keep changing

Study 1 · AI-peer project

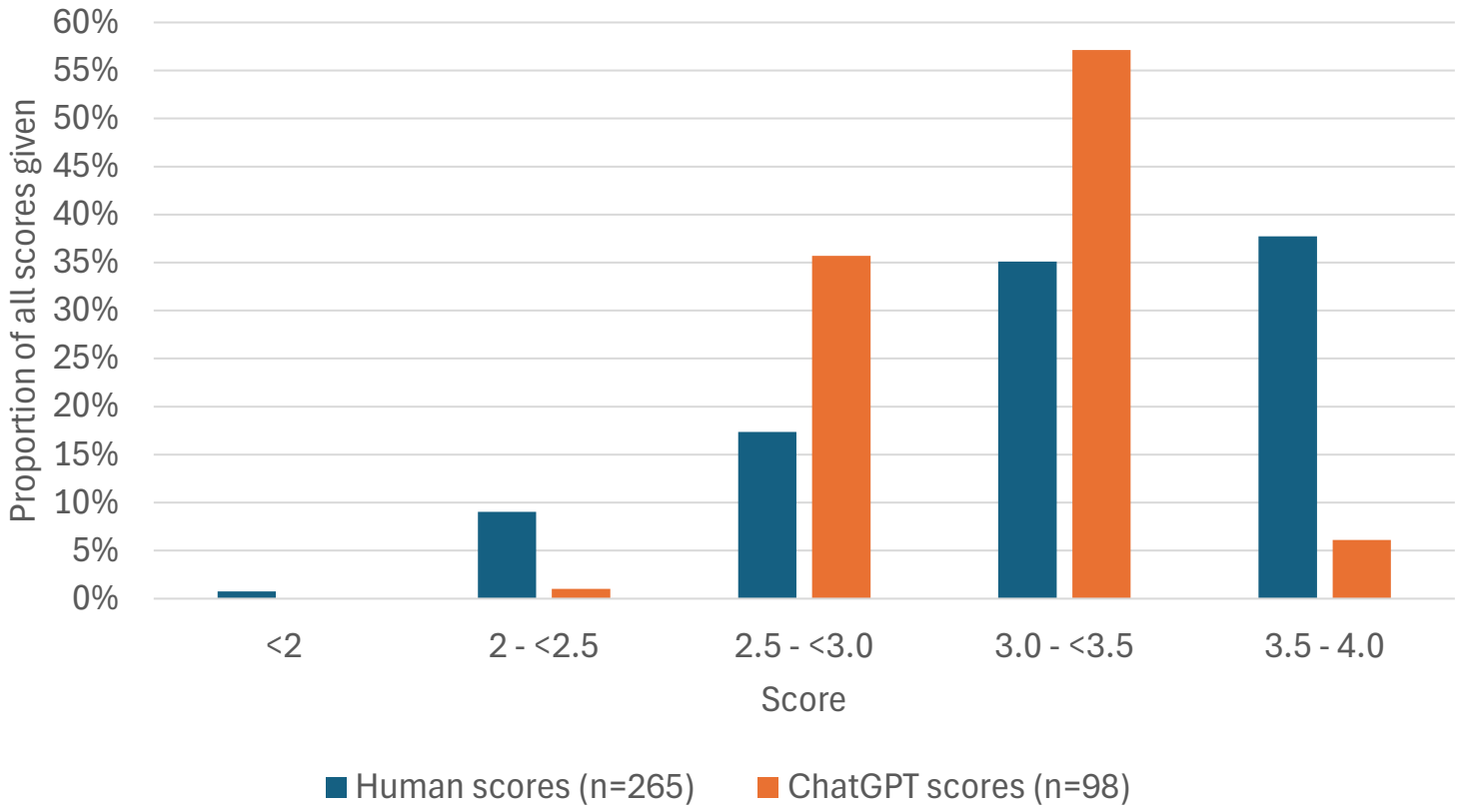
ChatGPT-4o, ChatGPT-4o mini, Gemini 2.0 Flash

Study 2 · AI-peer project

ChatGPT 5.2

LLMs are developing rapidly, and their capabilities continue to improve. As a result, both the performance of these models and their potential role in research assessment are likely to change over time.

Narrower, more compressed score distributions



- Human reviewers assigned substantially more scores in the top category (3.5–4.0: 37.7% vs 6.1%).
- ChatGPT rarely assigned very low scores either.
- LLM assessments avoid both extremes of the scale — especially the highest scores that human reviewers are more willing to give.

Example of LLM review report of an article

> A solid, internationally recognisable baseline audit with clear applied relevance and a useful empirical dataset, but limited by uneven sampling across procedures, largely descriptive analysis without uncertainty, incomplete specification of mass-estimation methods, and comparability issues for the COVID-era PPE component. With strengthened statistical treatment (confidence intervals, inter-rater reliability), balanced sampling, transparent item-weight protocols, and a more defensible extrapolation (with sensitivity analysis and reconciliation of reported means), the work could approach internationally excellent standards. As presented, it is best characterised as a credible and useful baseline study rather than a world-leading contribution. Overall score: 2.7

RESULTS

Example of LLM review report of an article

Rigour report: This is a competently executed descriptive, multi-country survey report with several methodological strengths: a clearly articulated aim (programme-level mapping), ethical approval, systematic instrument development (Delphi-style consensus with 11 experts).

However, the rigour is constrained by design and implementation choices that limit internal and external validity. Most importantly: (i) the sampling frame is essentially ADEE membership generating selection bias and uneven geographical coverage; (ii) the response rate (49.3%) is acceptable for institutional surveys but still leaves substantial nonresponse bias risk, and the paper does not report any nonresponse analysis; (iii) the study relies on single-informant institutional self-report without independent verification; (iv) the questionnaire is English-only, which plausibly affects comprehension and item nonresponse in some contexts; (v) “Not applicable” responses appear conflated with “prefer not to say / unknown creating ambiguity in interpretation

LLM text: coherent and meaningful

Well-trained on evaluative language
The models seem well trained on similar evaluative texts — research quality, standard review criteria, and the strengths and weaknesses of different scientific methods

Focused on broad quality characteristics
Generalisability, methodological novelty, theoretical contribution, and reproducibility — general markers rather than field-specific ones

Common limitations flagged
Most often: low response rates and limited generalisability

LLM text characteristics

Highly standardised, minimal variation
Review reports were coherent but highly standardised in wording and content — with minimal variation across repeated runs.

Echoes the guidelines, but stays general
Frequently reproduced language straight from the review guidelines, but often in a manner that remained general rather than specific to the article.

Format and structure differ substantially

LLMs

- Highly standardised format, directly following the prompt
- Always separate assessments of rigour, originality, and significance
- Comments referred to specific aspects of study design, analysis, or reporting to justify scores

Human reviewers

- Much more heterogeneous — from detailed assessments to a single line.
- Often emphasised topic relevance or interest without systematic methodological discussion.
- Numerical ratings sometimes came with little or no explanation.

Tone and voice: hedged vs. uniform

LLMs

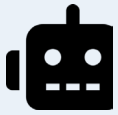
- Uniform confidence, every time
- Impersonal style across all papers
- Expressing evaluations with equal certainty regardless of topic

Human reviewers

- Often wrote in the first person
- Qualified their assessments when a paper fell outside their expertise.
- “not my area of expertise”

LLM provides detailed, consistently structured evaluations; human reviewers contribute disciplinary knowledge and contextual judgement, in a less standardised manner.

Generic evaluation vs. contextual judgement



LLM focus

- Broad characteristics: clarity, generalisability, methodological novelty, theoretical contribution, reproducibility.
- Strictly kept to the three prompt criteria — originality, significance, rigour.
- Neutral to topic: all research areas presented as important and interesting.
- Evaluative terms (e.g. “world-leading”) sometimes used without clear justification.

Human focus

- More specific: concrete methodological issues, contextual detail, domain-specific concerns.
- Sometimes commented beyond the criteria: structure, writing quality, accessibility to lay readers, bibliography.
- Reflect the values and interests of their academic field — topics are not evaluated as equally important.

LLM reports: long, wordy, and repetitive

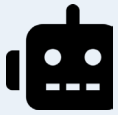
LLM

- Substantially longer than the human-written departmental review summaries.
- Texts were often wordy — using many words to make a single point, repeated with different formulations – hard to read
- When fed the same output repeatedly, reports stayed similar, though never identical, and sometimes with different scores
- For example, one 180-word abstract was assessed by an 800-word ChatGPT review.

Human reviewers

- Much shorter

Originality



LLMs

- Large-scale data or broad methodological scope acknowledged as a major strength
- Framing predominantly methodological rather than field-specific, with less explicit references to the actual research field
- Does not capture disciplinary novelty but reflect a more generic appraisal of conceptual and methodological newness

Human reviewers

- Assessment of significance of an advance that is difficult to assess without deep domain expertise
- Human reviewers captured forms of significance LLM missed

IMPLICATIONS

Possible consequences of substituting peer review with LLM

Score compression
Matters most when identifying top and bottom of a distribution apart — for example, separating genuinely world-leading work from merely competent work.

Domain blindness
Because the LLM's judgments stay generic and engage less with the actual research field, substitution risks underrecognising the field-specific novelty a domain expert would catch

Reproducibility and validity
Repeated runs on the same input give similar but not identical output — sometimes with a different score — a fairness concern if used for decisions. Moderate score correlation with peer reviews

Can LLMs be guided to review like humans?

Yunhan Yang, Judita Preiss and Henrik Karlstrøm

NIFU seminar August 13, 2026

Introduction

Using LLMs to understand what makes LLMs work or not as reviewers of scientific work

Four experiments, one common empirical core

- Three families of LLMs (Llama-3.1-8B, Falcon-3-10B, and Llama-3.1-405B) with a grid of parameter settings and variations in prompts generating reviews of 98 papers

Experiment 1: Guiding scores

To what extent does providing suggested review scores change LLM-generated reviews?

LLMs generated reviews without and with guidance scores

Each run also generated «evidence frames», providing examples of reasons to give a particular score

Experiment 1, cont.

Guiding changes LLM review scores but does not induce pure copying

- Similarity to human scores improved with guidance
- 31.9 % guide-copy rate
- Underestimates became more positive; overestimates needed more critical framing

Humans and LLM reviews provided different evidence

- Humans more focused on impact and significance claims, while LLMs emphasised questions of generalisability and methodology

Experiment 2: Predicting guiding scores

For out-of-sample applications, scores may not be available and must be predicted

Can predictions be based on pseudo-labels derived from other features?

Scores generated from bibliometric data from the OpenAlex database, based on the same field classification as the sample

- 52k health papers, reduced to ~3200 after filtering for full text availability

Experiment 2, cont.

Fairly low exact accuracy: 0.568 (ordinal regression)

High adjacent accuracy: 93 % of scores within one level

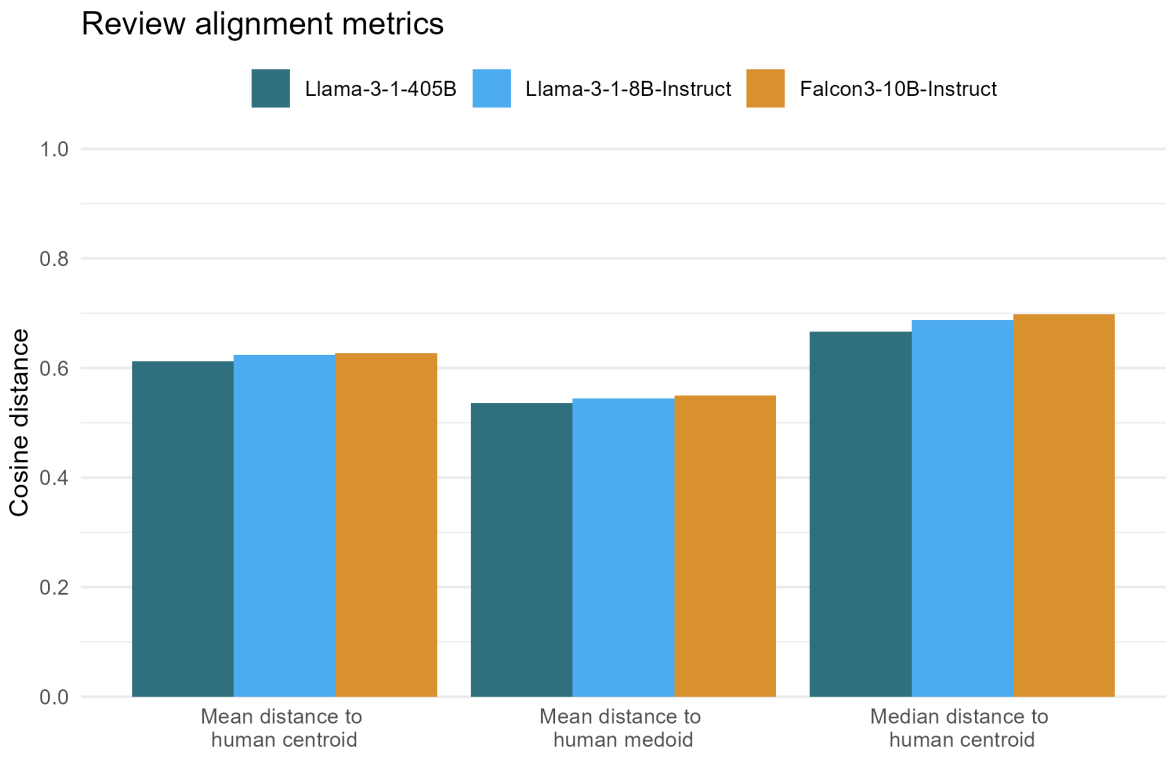
The result is promising but not solved: pseudo-score noise and lack of full texts limited data quality. Pseudo-scores help, but expert data remain essential

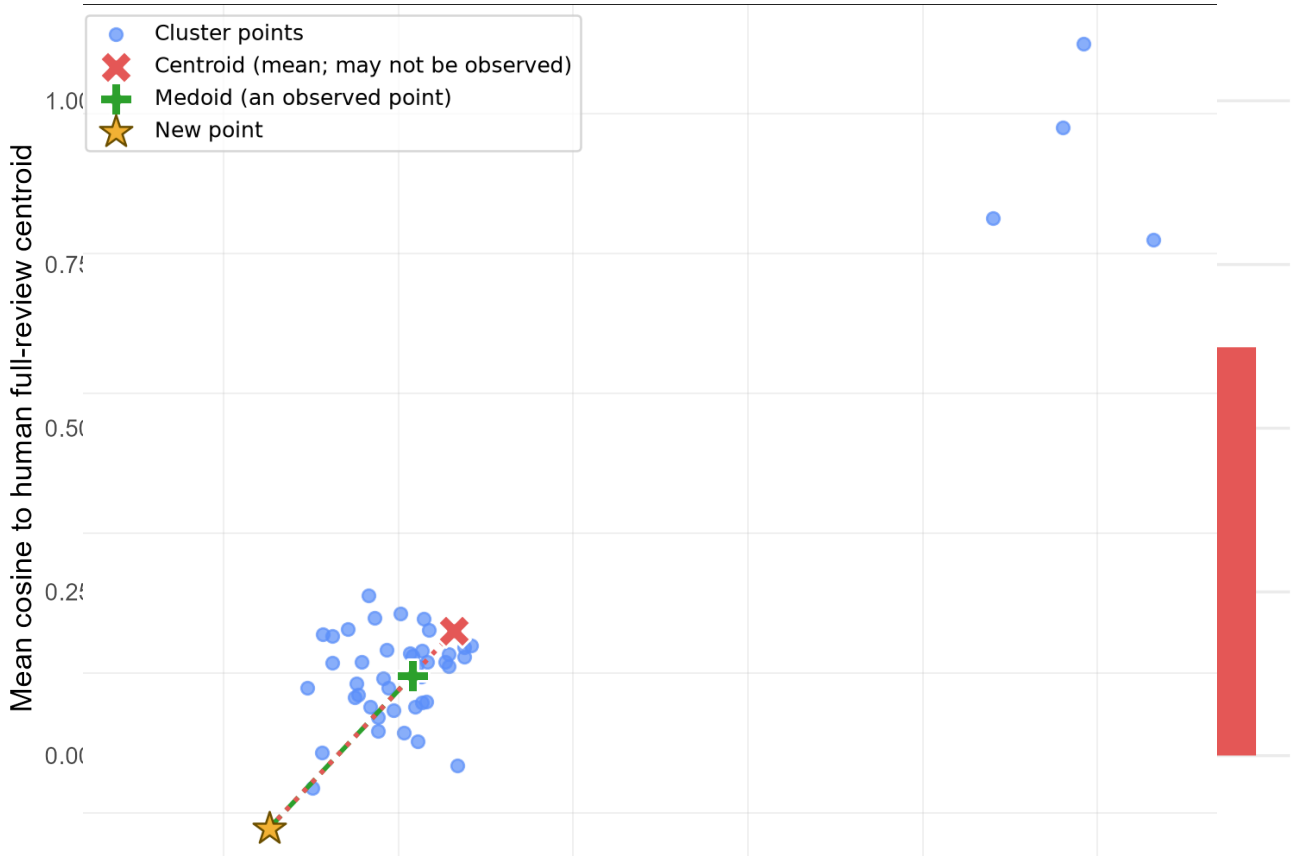
Experiment 3: Human-LLM review similarity

Using the internal geometric language representation of LLMs to assess similarity between human and LLM-generated reviews

Embeddings were retrieved from a separate model for human and LLM-generated reviews and various distance measures computed

Experiment 3, cont.





Experiment 4: LLMs as judges

LLM-generated reviews judged by a panel of LLMs with prompts for:

1) Adherence to REF-style criteria (on 0-3 scale)
2) Specificity of claims («generic», «mixed» or «paper-specific»)
3) Evaluative depth («descriptive», «limited», «analytical»)
4) Focal balance («strength only», «weakness only», «both»)
5) Extent of generic evaluative language («none», «limited», «substantial»)

Judgment passed by synthetic LLM panel of judges (GPT-5.6 Luna), validated by human and different LLM model (GPT-5.6 Terra) on subset

Experiment 4, cont.

High adherence in general (~90 % of scores are 2s), but varies across criteria

- Originality justified with reference to concrete features, rigour systematically lower scored but with high evaluative depth, significance weakest criterion

Later models more specific, more analytical and more balanced

More specific prompts more specific, more analytical and more balanced

Providing the full text increases paper-specific language but reduces balance

Significant amounts of generic language

AI Peer **Policy Brief**

How (not) to use Large Language Models to assess research

Presented by Liv Langfeldt

NIFU Nordic Institute for Studies in Innovation, Research and Education

The issue

Challenges in Research Assessment

Increasing manuscript volumes and reviewer fatigue strain traditional peer review systems.

Promise of LLMs

Large language models offer scalability and efficiency, potentially reducing workloads and speeding up assessments.

Implications

Using LLMs requires understanding impacts and ensuring alignment with scientific and societal goals.

What Current Studies Tell Us

Correlation with Expert Scores

LLM evaluations correlate moderately well with expert reviewer scores. Correlation may be higher than citation metrics' correlation with expert scores.

Middle range scores

LLMs tend to give upper-middle scores, compressing score distributions and limiting extreme evaluations.

→ Better at ranking than predicting individual scores

Probabilistic Nature of Scores

LLM scores reflect probabilistic predictions from training data. Can be made based on title and abstract alone.

LLMs develop rapidly ...

LLMs vs Peer Review

Peer review

1. Field specific expertise

- Assessments based on field standards/knowledge base
- Aim to advance the field/form quality notions
- Dynamic process

2. Reviewers selected based on expertise (and availability)

3. Level of expertise vary

LLMs

1. Trained on enormous corpora

- Mimic/guess likely assessments

2. Spead/efficiency

3. One “reviewer” can review all fields

How come scores correlate?

LLM biases

Availability bias

- Type of text and databases that dominate the training corpus
- Material that is easy to crawl and index

Averaging bias

- Heterogenous discourse → concise user-friendly text
- Low-frequency views ignored

Kerche FW, Zook M & Graham M (2026).
The silicon gaze: A typology of biases and inequality in LLMs through the lens of place.
Platforms & Society https://doi.org/10.1177/29768624251408919

Potential implications of using LLM reviews?

Peer review

An integrated activity in the scientific community

Develop research fields and their quality notions

LLM reviews

Generic criteria and assessments

Unlimited review capacity, with high speed

→ Possibility of constant monitoring and censorship
→ More fixed standards across fields?
→ Bias against novelty and small field?

Summary: Concerns About LLM-Based Assessment

Availability Bias in LLMs

LLMs favor dominant scientific traditions due to training on heavily represented data.

Averaging Bias Effect

May mask disagreement and disadvantage minority perspectives and small/emerging fields.

Unintended Incentive Effects

Researchers may adapt work to maximize AI scores, reinforcing conformity and discouraging innovation.

Illusion of Objectivity

May give impression that research quality is fixed and measurable, ignoring complexity and context.

Five Key Recommendations

Support, not replace

LLMs may assist but not replace expert judgement. Keep human experts accountable.

Check copyright

Must comply with copyright and data confidentiality.

Define evaluative context

LLMs are more useful for ranking comparable outputs than assigning individual scores.

Check for biases

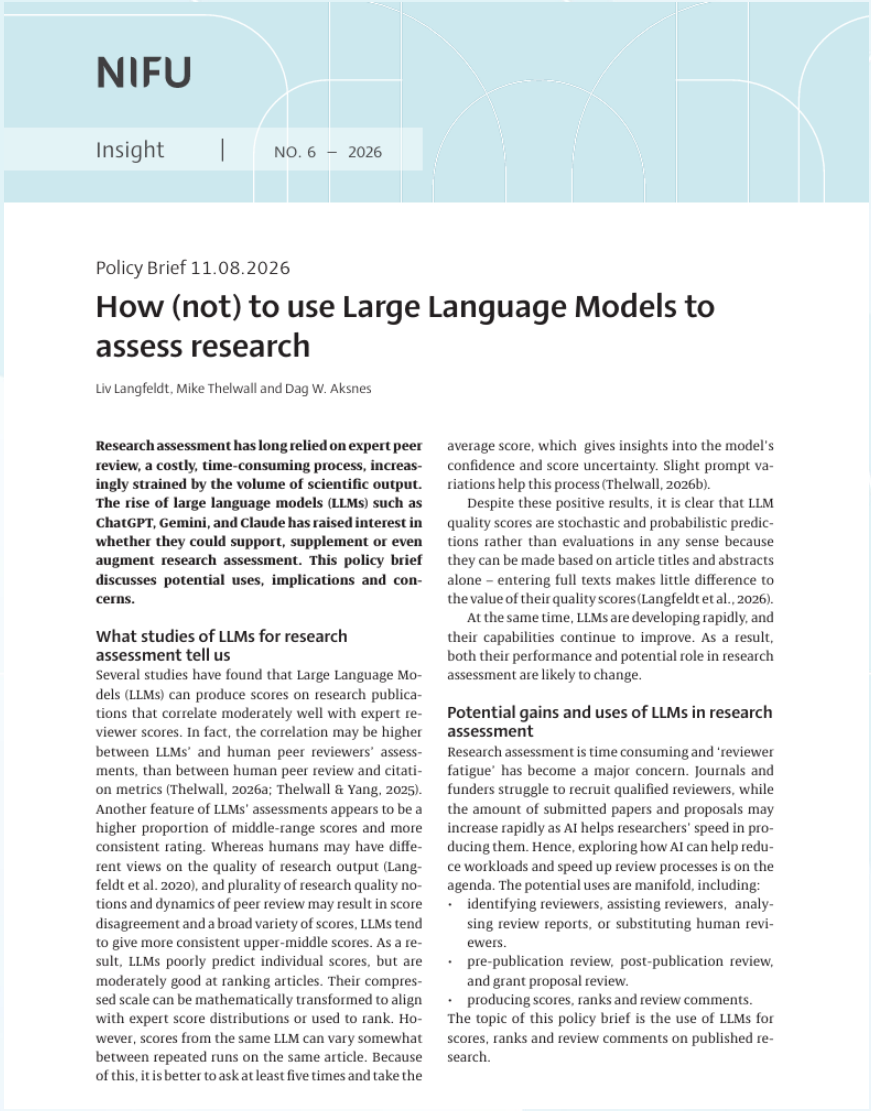
Incorporate bias detection and mitigation through model comparisons, prompt tweaks, and expert review.

Consider organizational goals and broader societal concerns

Consider potential effects on research diversity, trust, incentives, and policy alignment.

NIFU Nordic Institute for Studies in Innovation, Research and Education

Thank you!



NIFU

Insight | NO. 6 – 2026

Policy Brief 11.08.2026

How (not) to use Large Language Models to assess research

Liv Langfeldt, Mike Thelwall and Dag W. Aksnes

Research assessment has long relied on expert peer review, a costly, time-consuming process, increasingly strained by the volume of scientific output. The rise of large language models (LLMs) such as ChatGPT, Gemini, and Claude has raised interest in whether they could support, supplement or even augment research assessment. This policy brief discusses potential uses, implications and concerns.

What studies of LLMs for research assessment tell us

Several studies have found that Large Language Models (LLMs) can produce scores on research publications that correlate moderately well with expert reviewer scores. In fact, the correlation may be higher between LLMs' and human peer reviewers' assessments, than between human peer review and citation metrics (Thelwall, 2026a; Thelwall & Yang, 2025). Another feature of LLMs' assessments appears to be a higher proportion of middle-range scores and more consistent rating. Whereas humans may have different views on the quality of research output (Langfeldt et al. 2020), and plurality of research quality notions and dynamics of peer review may result in score disagreement and a broad variety of scores, LLMs tend to give more consistent upper-middle scores. As a result, LLMs poorly predict individual scores, but are moderately good at ranking articles. Their compressed scale can be mathematically transformed to align with expert score distributions or used to rank. However, scores from the same LLM can vary somewhat between repeated runs on the same article. Because of this, it is better to ask at least five times and take the average score, which gives insights into the model's confidence and score uncertainty. Slight prompt variations help this process (Thelwall, 2026b).

Despite these positive results, it is clear that LLM quality scores are stochastic and probabilistic predictions rather than evaluations in any sense because they can be made based on article titles and abstracts alone – entering full texts makes little difference to the value of their quality scores (Langfeldt et al., 2026).

At the same time, LLMs are developing rapidly, and their capabilities continue to improve. As a result, both their performance and potential role in research assessment are likely to change.

Potential gains and uses of LLMs in research assessment

Research assessment is time consuming and 'reviewer fatigue' has become a major concern. Journals and funders struggle to recruit qualified reviewers, while the amount of submitted papers and proposals may increase rapidly as AI helps researchers' speed in producing them. Hence, exploring how AI can help reduce workloads and speed up review processes is on the agenda. The potential uses are manifold, including:

- identifying reviewers, assisting reviewers, analysing review reports, or substituting human reviewers.
- pre-publication review, post-publication review, and grant proposal review.
- producing scores, ranks and review comments.

The topic of this policy brief is the use of LLMs for scores, ranks and review comments on published research.

AI Peer Policy Brief

Liv Langfeldt, Mike Thelwall and Dag W. Aksnes (2026) **How (not) to use Large Language Models to assess research.** NIFU Innsikt 2026:6
https://hdl.handle.net/11250/5555680

liv.langfeldt@nifu.no

www.nifu.no

https://sites.google.com/sheffield.ac.uk/peer-ai/home

NIFU Nordic Institute for Studies in Innovation, Research and Education

Can AI assess research?

NIFU 13th August 2026, 13.30 -15.00

Agenda

1. **Welcome** by Fredrik Piro, Head of Research, NIFU
2. **How can LLMs best be used for the post-publication research quality assessment of journal articles?**
   Mike Thelwall, University of Sheffield
3. **Differences between human and LLM review reports**
   Dag W. Aksnes, NIFU
4. **Can LLMs be guided to review like humans?**
   Henrik Karlstrøm, NIFU
5. **Policy brief 'How (not) to use Large Language Models to assess research'**
   Liv Langfeldt, NIFU
6. **Open discussion/Q&A**

The AI Peer project
https://sites.google.com/sheffield.ac.uk/peer-ai/home

Policy Brief
NIFU Innsikt 2026:6
https://hdl.handle.net/11250/5555680

NIFU Nordic Institute for Studies in Innovation, Research and Education