# Can smaller large language models evaluate research quality?

**Mike Thelwall***
School of Journalism, Media and Communication,
University of Sheffield, 2 Whitham Rd, Sheffield S102AH, Sheffield, UK
e-mail: *m.a.thelwall@sheffield.ac.uk (corresponding author)
ORCID ID: 0000-0001-6065-205x

## ABSTRACT

*Academic librarians often construct bibliometric indicators to support research evaluation. Traditionally, these have been citation-based, but AI alternatives have recently emerged. Although both Google Gemini (1.5 Flash) and ChatGPT (4o and 4o-mini) provide research quality evaluation scores that correlate positively with expert scores in nearly all fields, and more strongly than citations in most, it is not known whether this holds for smaller Large Language Models (LLMs). In response, this article assesses Google's Gemma-3-27b-it, a downloadable LLM (60 GB). Results for 104,187 articles show that Gemma-3-27b-it scores correlate positively with an expert research quality score proxy for all 34 Units of Assessment (broad fields) from the UK Research Excellence Framework 2021. The Gemma-3-27b-it correlations have 83.8% of the strength of ChatGPT 4o and 94.7% of the strength of ChatGPT 4o-mini correlations. Unlike the two larger LLMs, the Gemma-3-27b-it correlations do not increase substantially when scores are averaged across five repetitions, its scores tend to be lower, and its reports are relatively uniform in style. Overall, the results show that research quality score estimation can be conducted by offline LLMs, so this capability is not an emergent property of only the largest LLMs. Moreover, score improvement through repetition is not a universal feature of LLMs. In conclusion, although the largest LLMs still have the highest research evaluation score estimation capability, smaller ones can also be used for this task, which can be helpful for cost saving or when secure offline processing is required.*

***Keywords**: Scientometrics; Large Language Models; Gemma; Open weights LLMs.*

## INTRODUCTION

Research evaluation is a common and important task for academics and managers, often supported by citation-based indicators (Hicks et al., 2015; Moed, 2005; Mukherjee, 2022). Academic librarians are frequently required to create or interpret these indicators, either as their primary responsibility or as one of their duties in a broader role. With the increasingly widespread use of Artificial Intelligence (AI) in research (Mohammadi et al., 2025), it is important to assess whether AI can save expert time by supporting the research evaluation process. This can highlight the value of AI for this task and provide broader insights into its usefulness for other information-related tasks performed by librarians. In this context, ChatGPT has demonstrated significant general value for academic librarians due to its ability to summarise documents, write literature reviews, answer questions, and search for information online. However, this article focuses solely on research evaluation.

ChatGPT research quality score estimates for journal articles are recent alternatives to citations as quantitative indicators to support evaluations (Kousha & Thelwall, 2025). Their value lies in their positive correlation with expert judgement in all or nearly all fields, and at

a slightly higher rate than citation-based indicators (Thelwall, 2025a; 2025b; 2025c). Despite some systematic biases or disparities (Thelwall & Kurt, 2025), this property means they are useful when expert judgement is unavailable, such as in areas outside the assessor's expertise, as a cross-check for bias, or for evaluations where assessment expertise is unavailable or too costly for the value of the task (Thelwall, 2025d). Librarians supporting researchers can therefore consider using ChatGPT's quality judgements instead of citation-based indicators in situations where citations are inappropriate, such as for recently published articles or for arts and humanities fields where citation counts have little relevance.

While a positive correlation with expert judgement has been established for three of the largest Large Language Models (LLMs) in 2025, ChatGPT 4o, ChatGPT 4o-mini, and Google Gemini Flash 1.5 (Thelwall, 2025a; 2025c), these are all cloud-based services and may be too expensive or insufficiently private for some research evaluation purposes (Nowak et al., 2025). Moreover, cloud-based services can be withdrawn, updated, or made more costly, so research evaluation procedures may not be able to rely on them. Therefore, there is a need to test whether any smaller "open weights" LLMs (Sowe et al., 2024) that can be downloaded and used offline are capable of estimating research quality. If these also work, academic librarians could use them to evaluate private documents, such as drafts in preparation, manuscripts under review by a journal, or copyright-protected published articles, which could not be uploaded to public LLMs.

From a broader perspective, further evidence of the capability of smaller LLMs that can run offline in private systems could encourage their use in wider contexts. One example is support for the peer review process. This must be private, as reviewers undertake not to share the reviewed work and therefore should not upload it to a public LLM. They could still use smaller private LLMs to provide advice, and some journals and conferences already offer this facility automatically to reviewers (Thakkar et al., 2025). The same capability is also used by at least one funder to help evaluate funding proposals (Carbonell Cortés et al., 2024). Thus, evidence of smaller LLMs' ability to evaluate the quality of academic work would provide further support for their use in these tasks.

## LITERATURE REVIEW

LLMs are a type of artificial intelligence with a specific design (architecture) and are trained on an enormous quantity of text. This training gives generative LLMs the capability to write grammatically correct and usually meaningful language in response to a user prompt. With an additional instruction tuning stage, they can also respond with reasonable answers to a wide variety of user requests (Chung et al., 2024). Multimodal LLMs are trained on images as well as text and can respond to prompts such as, "Generate an appropriate caption for this image" (Qi et al., 2023). Another relevant concept is the "mixture of experts" (Jacobs et al., 1991), the idea that a single AI system can include multiple subsystems specialising in different tasks. LLMs can therefore produce substantially different results for the same input if it is routed to a different "expert" sublayer (Shazeer et al., 2017). LLMs are inherently language-agnostic but are typically trained on multiple languages and can perform services such as translation as well as responding to prompts in multiple languages (Huang et al., 2024).

One unusual phenomenon is that some LLM capabilities do not exist in smaller versions but only appear when the size is large enough; these are called emergent properties (Wei et al.,

2022). The opposite is a predictable, scalable property that may be weak in smaller models but becomes stronger as model size increases. Because of the potential for emergent properties, large LLM capabilities cannot be guaranteed to exist in smaller LLMs without testing. Thus, it is not clear whether the research quality scoring capability of ChatGPT and Gemini also exists in smaller LLMs.

The magnitude of an LLM can be judged by its file size or the number of parameters when saved. LLMs can be made smaller through quantisation, which retains the number of parameters but saves them at a reduced level of accuracy (Xiao et al., 2023). Overall, the capability of an LLM depends on its number of parameters, overall design (architecture), and the precision with which the parameters are stored. LLMs are available for download at various sizes, from a few gigabytes to hundreds of gigabytes, on the huggingface.co platform.

Although most investigations of LLMs for research quality estimation have used ChatGPT or Gemini, as mentioned in the introduction, one previous study has used locally run smaller LLMs. It used different versions of Llama3 and Qwen2.5, comparing them with ChatGPT 4o-mini. The task was a three-class proxy quality rating for biomedical research. Zero-shot and few-shot were tried (and fine-tuning for non-generative models). Qwen 2.5-72b and ChatGPT 4o-mini performed best, but precision and recall were reported rather than correlations, so it is not possible to judge whether the smaller LLMs had the capability to rank articles for research quality. Nevertheless, the highest Cohen's kappa reported was only 0.059 (for zero-shot), suggesting a very weak ability for this (Wu et al., 2025). Positive results have been obtained from smaller LLMs for the different task of pre-publication conference paper reviewing, however (Zhou, Chen, & Yu, 2024).

## METHODS

As discussed above, the capability of smaller LLMs has not been directly assessed for research quality scores, nor has it been evaluated outside of biomedicine. The objective of the current study is to address this gap with a comprehensive study across all scientific fields, directly investigating research quality scores through the following research questions. The second question, concerning the need for repetitions, is relevant because previous research has shown that averaging multiple independent scores for the same article from ChatGPT or Gemini yields a higher correlation with expert scores than individual scores. This needs to be tested for smaller LLMs, as repetition increases the cost of scoring and should be avoided if unnecessary.

- RQ1: Can smaller (i.e., downloadable open weights) LLMs provide research quality scores that correlate positively with expert judgement across all fields of science?
- RQ2: Does averaging LLM scores across multiple repetitions increase the correlation with expert judgement?

For RQ1, the research design involved scoring a large dataset of journal articles from all scientific fields using the smaller LLM Gemma and then correlating these scores with an indicator of the expert quality scores for those articles. For RQ2, the average correlations with individual scores were compared with the correlations obtained from average scores across five repetitions.

### Journal article dataset

As in all previous science-wide evaluations of LLMs, this article uses journal articles without short abstracts submitted to the UK's Research Excellence Framework (REF) 2021 as the

dataset for analysis. This is a large set of primary research articles (excluding reviews) published between 2014 and 2020, selected by researchers in UK higher education institutions as their best one to five outputs during the period. Each output was given an individual score of 1* (nationally relevant), 2* (internationally relevant), 3* (internationally excellent), or 4* (world leading) by two senior researchers from the relevant broad field during a year. This is the largest systematic research quality evaluation ever conducted and is a serious process, as the results determine the UK's annual block grant for research, which totals about £14 billion over the lifetime of REF2021. Unfortunately, the scores for individual journal articles are not released, but average scores for sets of articles are public. These sets are approximately department-sized (the terminology used here). REF 2021 was organised into 34 broad disciplines, called Units of Assessment (UoAs), and each university could submit the work of one group of researchers (two in some circumstances) to each UoA.

The public results for the REF are the average scores for each university and UoA (or group and UoA when more than one submission from a university to a UoA). Since the average scores vary substantially between universities, these averages are reasonable proxies for the individual article scores. Thus, the dataset analysed here consists of REF2021 journal articles and the associated average departmental research quality scores. For each UoA, articles with the 10% shortest abstracts were removed. The 10% threshold was determined heuristically to exclude articles without abstracts or those that were short-form contributions (e.g., research notes) not equivalent to standard journal articles. Since the same dataset has been used previously, the correlations for those studies are directly comparable to those reported here.

For research quality scoring, each article is represented by its title and abstract rather than its full text. This is for two reasons. First, ChatGPT's scores from titles and abstracts show higher correlations with expert judgement than its scores from full text. Second, not all article full texts are available online without charge.

**Research quality scores from Gemma**

Google Gemma-3-27b-it is a multimodal, multilingual, instructiontuned open weights LLM that can be downloaded from huggingface.co (https://huggingface.co/google/gemma-3-27b-it) with a combined file size of 5 9.9 GB in safetensors format and 27 billion parameters. Its total size may be misleading because its image processing capabilities are not used here. Gemma is described by Google as a family of "lightweight" LLMs that it shares freely with the research community. The current version, Gemma 3, has a 128k token context window, which easily allows it to process article titles and abstracts. It is available in four sizes, based on the number of parameters, with the smallest having 1 billion and the largest 27 billion (as used here). The models were released in March 2025 (https://developers.googleblog.com/en/introducing-gemma3/).

All articles in the dataset were submitted to Gemma 3 for evaluation with the same system instructions as previously used for ChatGPT and Gemini. These instructions (see: Thelwall, 2025b) define the research quality evaluation task. The system instructions are slight paraphrases of those given to the REF2021 expert reviewers. There were four sets of expert reviewer instructions and, hence, four corresponding system instructions. These are for the groupings of UoAs into main panels A (health and life sciences, UoAs 1–6), B (engineering and physical sciences, UoAs 7–12), C (social sciences, UoAs 13–24), and D (arts and humanities, UoAs 25–34).

Each set of system instructions explains that the task is to evaluate the quality of a journal article, defines the four different star levels, discusses the factors that might be considered

when assigning a score, and defines research quality as comprising rigour, originality, and significance. These are the most common dimensions for research quality definitions (Langfeldt et al., 2020).

For each evaluation, Gemma was provided with the appropriate system instructions for the article (based on its UoA), then the user prompt "Score this:" followed by the article title, a new line, the word "Abstract", another new line, and the abstract on a single line. Each score request was submitted in a separate session to avoid the score from one article influencing the score for another. This was repeated five times (non-consecutively) for each article, and the arithmetic mean of the five scores was used as the final score. Although Gemma is an open-weights LLM, this process was conducted through the free Google Gemma API (https://ai.google.dev/gemma/docs/core/gemma_on_gemini_api).

The Gemma output from the above prompts is typically a report evaluating the article. Within the report, there is almost always a score on the specified scale. This was extracted by a programme for the purpose (available at https://github.com/MikeThelwall/Webometric_Analyst). The recommended score was always extracted, except that if Gemma gave a fractional score (e.g., 2.67 and then rounded it to an exact star level, 3* in this case), the unrounded number was used.

**Correlations**

The LLM scores were correlated with the departmental average research quality scores separately for each UoA, as the average scores vary between UoAs and UoA-level results may reveal field differences in LLM capability. Spearman correlations were used because the primary use of the LLM scores is to rank articles rather than to assign a specific score. For example, the quantitative indicators provided to some REF2021 evaluators were percentile ranks rather than score predictions. Bootstrapping was used to calculate the Spearman correlation confidence intervals, as there is no formula for this.

## RESULTS

Overall, Gemma rarely used the lowest score (1*) but still tended to give lower scores than the REF experts overall. Gemma also tended to avoid giving scores higher than 3* for social sciences, arts, and humanities fields. Table 1 shows descriptive statistics for the dataset analysed and the Gemma scores. Totals and averages are from UoAs, not articles (i.e., including double counting for articles in multiple UoAs).

The Gemma scores correlated positively with the departmental average scores in all fields, and the correlations were statistically significantly different from zero in 30 of the 34 UoAs, except for smaller UoAs (426 or fewer articles). Figure 1 shows Spearman correlations and 95% bootstrapped confidence intervals for Gemma research quality scores against departmental average REF scores, used as a proxy for individual article scores, by UoA. Thus, Gemma demonstrates a universal or near-universal ability to estimate research quality scores, although this ability is weak in most fields (below 0.3) or moderate.

Table 1: Descriptive statistics for the dataset analysed and the Gemma scores.

| UoA | Articles | REF mean | Gemma minimum | Gemma maximum | Gemma mean |
|---|---|---|---|---|---|
| 1. Clinical Medicine | 9604 | 3.28 | 2 | 4 | 2.97 |
| 2. Public Health, Health Services and Primary Care | 3822 | 3.34 | 2 | 4 | 2.94 |
| 3. Allied Health Professions, Dentistry, Nursing & Pharmacy | 9340 | 3.16 | 2 | 4 | 2.77 |
| 4. Psychology, Psychiatry and Neuroscience | 7876 | 3.11 | 2 | 4 | 2.87 |
| 5. Biological Sciences | 6124 | 3.32 | 2 | 4 | 2.99 |
| 6. Agriculture, Food and Veterinary Sciences | 3070 | 3.12 | 2 | 4 | 2.84 |
| 7. Earth Systems and Environmental Sciences | 3378 | 3.34 | 2 | 4 | 2.92 |
| 8. Chemistry | 2325 | 3.40 | 2 | 4 | 2.90 |
| 9. Physics | 3634 | 3.39 | 2 | 4 | 2.95 |
| 10. Mathematical Sciences | 2965 | 3.39 | 2 | 4 | 2.92 |
| 11. Computer Science and Informatics | 4122 | 3.20 | 2 | 4 | 2.56 |
| 12. Engineering | 14358 | 3.24 | 1 | 4 | 2.59 |
| 13. Architecture, Built Environment and Planning | 2239 | 3.14 | 2 | 3 | 2.41 |
| 14. Geography and Environmental Studies | 2788 | 3.11 | 2 | 3.2 | 2.79 |
| 15. Archaeology | 426 | 3.13 | 2 | 4 | 2.82 |
| 16. Economics and Econometrics | 702 | 3.23 | 2 | 4 | 2.82 |
| 17. Business and Management Studies | 8939 | 3.05 | 2 | 3 | 2.55 |
| 18. Law | 1471 | 2.99 | 2 | 4 | 2.50 |
| 19. Politics and International Studies | 2077 | 3.01 | 2 | 3 | 2.73 |
| 20. Social Work and Social Policy | 2724 | 2.96 | 2 | 3 | 2.49 |
| 21. Sociology | 1266 | 2.98 | 2 | 3 | 2.63 |
| 22. Anthropology and Development Studies | 750 | 2.99 | 2 | 4 | 2.75 |
| 23. Education | 2852 | 2.94 | 1.8 | 3 | 2.43 |
| 24. Sport and Exercise Sciences, Leisure and Tourism | 2423 | 3.13 | 2 | 3.6 | 2.48 |
| 25. Area Studies | 383 | 3.07 | 2 | 3 | 2.52 |
| 26. Modern Languages and Linguistics | 719 | 3.03 | 2 | 3 | 2.50 |
| 27. English Language and Literature | 516 | 2.92 | 2 | 3 | 2.45 |
| 28. History | 673 | 2.84 | 2 | 3 | 2.70 |
| 29. Classics | 54 | 3.05 | 2 | 3 | 2.54 |
| 30. Philosophy | 419 | 3.04 | 2 | 3 | 2.71 |
| 31. Theology and Religious Studies | 135 | 2.71 | 2 | 3 | 2.48 |
| 32. Art and Design: History, Practice and Theory | 811 | 2.89 | 2 | 3 | 2.25 |
| 33. Music, Drama, Dance, Performing Arts, Film & Screen Studies | 417 | 2.85 | 1 | 3 | 2.34 |
| 34. Comm, Cultural & Media Studies, Library & Information Management | 785 | 2.98 | 2 | 3 | 2.42 |
| **Total/Average** | 104,187 | 3.10 | - | - | 2.66 |

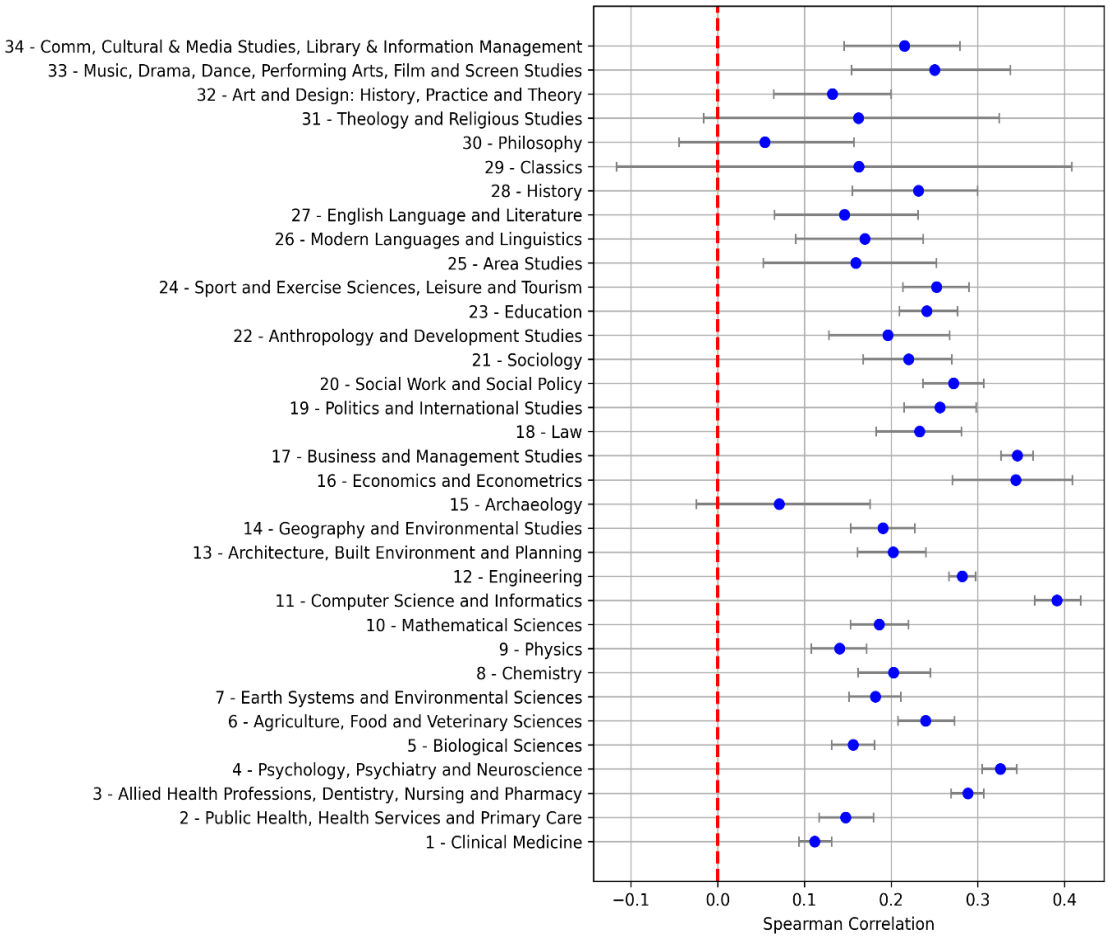


Figure 1. Spearman correlations and bootstrapped confidence intervals for Gemma research quality scores against departmental average REF scores

Averaging multiple iterations of Gemma makes little difference to the results. Figure 2 presents Spearman correlations between Gemma research quality scores and departmental average REF scores, used as a proxy for individual article scores, for different numbers of Gemma iterations averaged, by main panel. Main Panels A and C almost perfectly overlap, and the y-axis does not start at 0. Averaging five iterations increased the correlation by 2% (Main Panels A and C), 1% (Main Panel B), or 0.1% (Main Panel D) overall. The main reason for the lack of change was that the score rarely varied. In 95.7% of articles, all five scores were identical. In contrast, averaging five iterations increased the Gemma correlation by 8% overall, with only 45% of articles having identical scores all five times (author's calculation from the raw data).

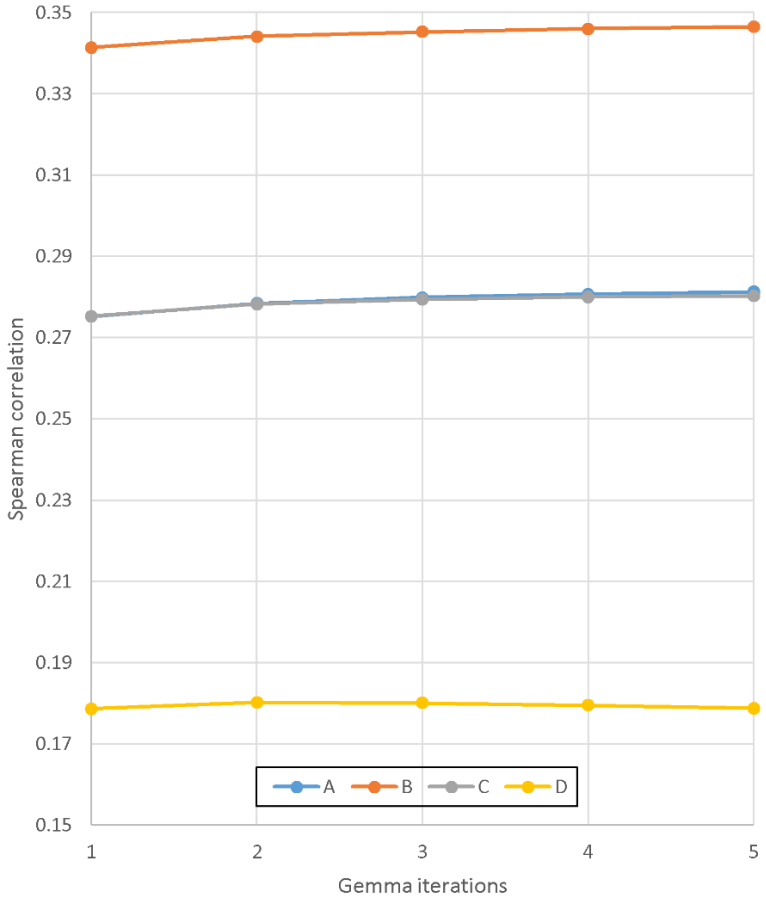


Figure 2: Spearman correlations for Gemma research quality scores against departmental average REF scores

## DISCUSSION

This study is limited by the use of a single UK-based dataset and departmental average quality scores rather than individual article quality scores. This likely dampens the observed correlations. Gemma may perform less effectively on outputs from other countries, especially if they are not in English. It may also be less able to estimate scores for different definitions of research quality. The results may differ for other LLMs, and some may be unable to estimate research quality scores at all. Finally, higher correlations might have been achieved obtained from other similarly sized LLMs or alternative strategies (e.g. few-shot learning, fine-tuning, or different system instructions).

### Comparison with prior research

Compared to ChatGPT 4o and 4o-mini (as reported in Thelwall, 2025b), the correlations with expert REF scores (again using the departmental average proxy) tend to be lower, although the difference is not large (Figure 3). Figure 3 presents Spearman correlations for ChatGPT 4o, ChatGPT 4o-mini, and Gemma-3-27b-it research quality scores against departmental average REF scores, used as a proxy for individual article scores, by UoA. The ChatGPT data are from Thelwall (2025b).

In a few cases, however, the Gemma correlations are higher (e.g., UoA 17). Using a sample size-weighted mean, the average Spearman correlations across all means are: ChatGPT 4o,

0.285; ChatGPT 4o-mini, 0.252; and Gemma-3-27b-it, 0.239. Thus, overall, the Gemma-3-27b-it correlations have 83.8% of the strength of ChatGPT 4o and 94.7% of the strength of ChatGPT 4o-mini. Surprisingly, despite the broadly similar correlations, the average score estimates were substantially higher for ChatGPT 4o (3.22) and ChatGPT 4o-mini (3.15) than for Gemma-3-27b-it (2.66).

Unfortunately, the correlations are not directly comparable to those previously published for Gemini Flash 1.5 (Thelwall, 2025a), as these were based on smaller and systematic samples: a maximum of 200 articles per UoA, in most cases taken from the highest and lowest scoring departments. Recall that the correlations are underestimates due to the use of a proxy research quality score. The results are also not directly comparable with the Cohen's kappa of the biomedical study because of the different metric, sampling method, and quality type (Wu et al., 2025). Nevertheless, the current results are more promising in the sense that they cover more fields and provide stronger evidence of an ability to rank articles for research quality.

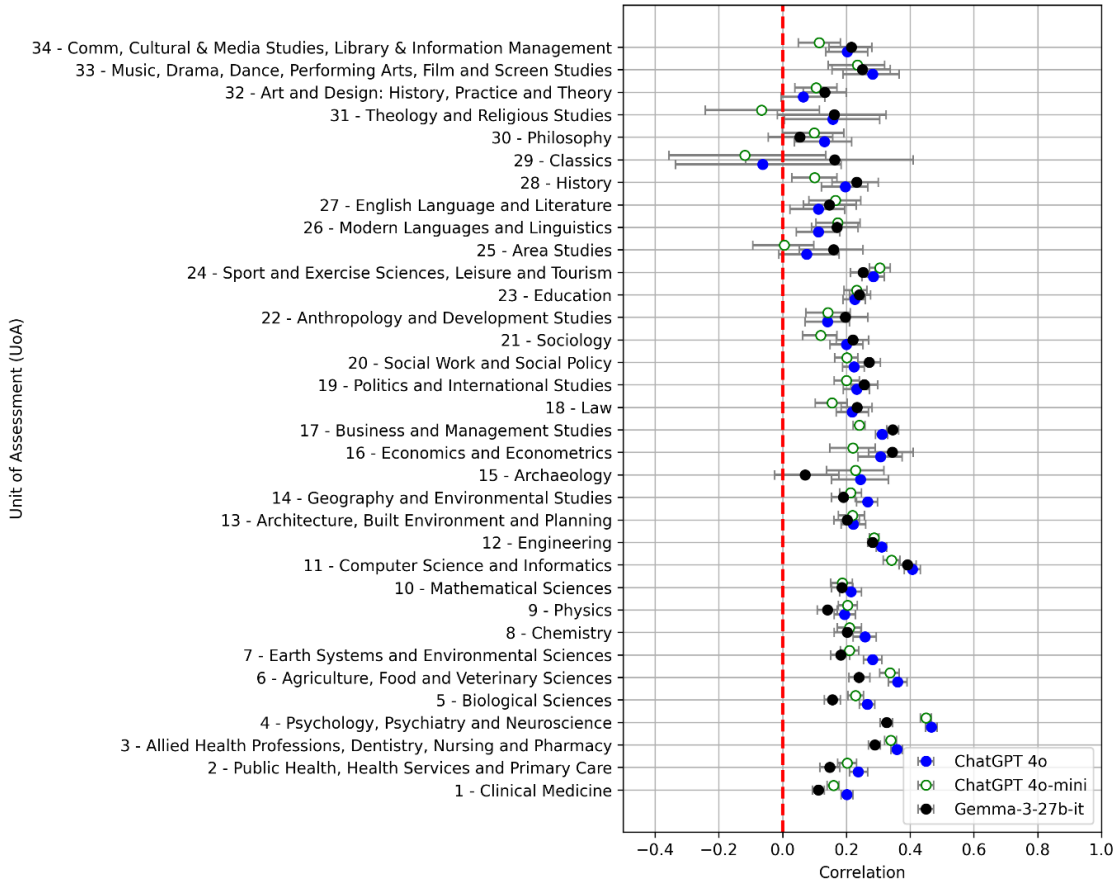


Figure 3: Spearman correlations for ChatGPT 4o, ChatGPT 4o-mini and Gemma-3-27b-it research quality scores against departmental average REF scores

**Report structures**

One clear difference between ChatGPT 4o/4o-mini and Gemma 3 is that Gemma's reports tend to follow a standard structured format, whereas ChatGPT 4o and 4o-mini use highly varied structures. For example, in the first set of reports for Main Panel D, the reports always began with a heading stating that the report was an assessment or evaluation of the article, including its title. The second paragraph was usually an overall score, typically followed by a heading "Justification" and then separate scores and justifications for Originality,

Significance, and Rigour (see Table 2). The sample report in the appendix illustrates the standard structure of these reports.

Table 2: The contents of the first three paragraphs of the first set of Gemma-3-27b-it reports for Main Panel D (n = 4966).

| Para. | Freq. | Content |
|---|---|---|
| 1 | 4,956 | "## Assessment of:", followed by title (with or without single/double quotes) |
| 1 | 10 | "## Evaluation of:", followed by title (with or without single/double quotes) |
| 2 | 3,446 | Overall score, with or without a descriptor. |
| 2 | 772 | Explanation that an assessment of the article abstract follows |
| 2 | 748 | Explanation that an assessment of the article follows |
| 3 | 196 | The phrase "Detailed assessment" |
| 3 | 2,840 | The phrase "Justification" |
| 3 | 331 | The phrase "Justification", followed by a justification |
| 3 | 273 | An originality score |
| 3 | 1,248 | Overall score, with or without a descriptor. |
| 3 | 57 | Evaluative summary of the paper. |
| 3 | 16 | The phrase "Rationale", followed by a justification |
| 3 | 5 | Here's a detailed breakdown of the assessment based on the provided criteria: |

Analysing the common headings, including emphasised phrases at the start of paragraphs, virtually all reports contained separate sections for rigour, originality and significance and an overall score. The vast majority also contained a section justifying the score and a conclusion or summary (Table 3).

Table 3: The most frequent headings or starts of paragraphs in Gemma-3-27b-it reports for Main Panel D (n = 24,830 reports)

| Heading/paragraph start | Frequency | Percentage |
|---|---|---|
| **3. Rigour… | 24,827 | 100.0% |
| **1. Originality… | 24,824 | 100.0% |
| **2. Significance… | 24,822 | 100.0% |
| **Overall Score… | 24,784 | 99.8% |
| **Justification… | 20,848 | 84.0% |
| **In conclusion… | 19,529 | 78.7% |
| ## Assessment of … | 12,556 | 50.6% |
| **Detailed Breakdown… | 8,327 | 33.5% |
| Here's an assessment of … | 7,319 | 29.5% |

On the few occasions when an article received different scores from Gemma, this was sometimes because Gemma reported an unrounded score in one report and only a rounded score in another (e.g., "Overall Score: 2.5* (Rounded to 3*)"). In very rare cases, different scores were given for the same criterion, as illustrated by the two sections below about significance, extracted from different reports for the same article. The reports concern a New England Journal of Medicine article presenting evidence from a clinical trial that a new treatment did not work. The second report includes additional negative points compared to the first. Generative LLMs produce text using a probability model, token by token, so the

justification is partly based on the score rather than the other way round. Thus, the additional criticisms are not the reason for the lower score; rather, the lower score could at least partly explain the additional criticism. Nevertheless, both the lower score and the additional criticism could be at least partly due to the model attaching more importance to effect size when reading the abstract. If so, the explanation given in the paragraph would help to understand the lower score. Unfortunately, it is not possible to test this.

> **2. Significance (3*):**
> [redacted summary of the study's significance] The study's results will likely influence clinical guidelines and future trial designs, potentially shifting focus towards alternative therapeutic strategies. The clear reporting of both primary and secondary outcomes, including adverse events, enhances its practical relevance. However, the lack of a positive finding limits its transformative impact on the field.
>
> **Significance (2*):**
> [redacted summary of the study's significance] This finding is important for guideline development and resource allocation. However, the effect size is small, and the confidence intervals are relatively wide, limiting the strength of the conclusions. The study's impact is more about refining understanding than dramatically altering it. While the findings will be noted and considered by clinicians, they are unlikely to fundamentally change the landscape of sepsis management.

The Gemma-3-27b-it reports are different from, and more standardised than, the reports from ChatGPT 4o for Main Panel D. For example, the common Gemma phrase, "Here's an assessment of" never appears in ChatGPT 4o reports, and the closest phrases starting, "Here is…" occur in less than 0.1% of ChatGPT 4o reports. At the simple level of characters or tokens, the common Gemma heading "1. Originality…" has many equivalents in ChatGPT 4o, including "- Originality:…", "### Originality:", "### 1. Originality:…", and "- Score for Originality:…". In terms of overall structure, most ChatGPT 4o reports (4435/4966) start with phrases or headings meaning "Evaluation of the article", but some reports start instead with a score, the article title, or a summary of the article.

## CONCLUSIONS

The results demonstrate, for the first time, that a downloadable open-weights LLM can possess a non-trivial ability to score academic journal articles for research quality, in the sense that its scores correlate positively with expert judgements. This capability is present in all or nearly all fields (not all correlations were statistically significantly different from zero). This implies that academic librarians can use downloadable LLMs to support decision-making in contexts where a quantitative indicator is required, such as situations where citation-based indicators are currently employed. They can also be used for important planned research evaluations, such as the REF, where a guarantee of performance and system availability would be required in advance. For a downloadable LLM, there is no risk that the necessary capability will disappear after a model upgrade. Although the correlations are weaker than those for ChatGPT 4o and ChatGPT 4o-mini, offline LLMs may be necessary for high-security contexts within closed systems or for cost savings in some situations. Switching to one of these models would result in a small overall reduction in performance, with 83.8% of the correlation strength of ChatGPT 4o and 94.7% of that of ChatGPT 4o-mini. Since the inputs are titles and abstracts rather than full texts, and LLMs are pattern-matching and processing algorithms, their scores are not evaluations but only guesses based on their world knowledge and the content of each title and abstract.

At a more theoretical level, the results show that research quality scoring is not an emergent property of only the largest current LLMs but is also present in the 27 billion parameter model used here. A 70 billion parameter version of the open-weights LLM Llama 3 is sometimes used in research (e.g., Wu et al., 2025), so 27 billion is relatively modest by comparison. Nevertheless, there may still be a lower threshold, in terms of the number of parameters, below which research quality scoring does not yield useful results. It may also not exist for other current LLMs of similar or larger size, but the main purpose of this paper was to show that the capability can exist in open-weights LLMs rather than to identify the best one.

The scores and structures of reports written by Gemma-3-27b-it are much less varied than those produced by ChatGPT 4o-mini. Thus, report variety may be an emergent property of the largest current LLMs. Their sizes or architectures may enable them to exploit their knowledge in more diverse ways, perhaps revealing different research evaluation perspectives within their knowledge, which allows their overall scores, after averaging multiple iterations, to be more robust.

Regarding broader lessons for librarians and information specialists about LLMs, this article has focused on a model that is much smaller than widely available web-based models such as ChatGPT, DeepSeek, and Claude. Smaller models are suitable for private or in-house systems when the large computing infrastructure required to run full-size models is unavailable or too costly. The results above add to the evidence that these smaller models can possess advanced capabilities that nearly match those of the largest models. Therefore, librarians and information specialists can consider recommending them for private document processing tasks. With these LLMs, such systems can incorporate more powerful processing capabilities than previously possible.

## ACKNOWLEDGEMENTS

This project was funded by the Economic and Social Research Council (ESRC), UK.

## CONFLICT OF INTEREST

The authors have no competing interests to declare in relation to the content of this article.

## REFERENCES

Carbonell Cortés, C., Parra-Rojas, C., Pérez-Lozano, A., Arcara, F., Vargas-Sánchez, S., Fernández-Montenegro, R., Casado-Marín, D., Rondelli, B. & López-Verdeguer, I. (2024). AI-assisted prescreening of biomedical research proposals: Ethical considerations and the pilot case of “la Caixa” Foundation. *Data & Policy*, *6*, e49. https://doi.org/10.1017/dap.2024.41.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y.-X., Wang, X.-Z., Dehghani, M., Brahma, S., Webson, A., Gu, S.-X. S., Dai, Z.-Y., Suzgun, M., Chen, X.-Y., Chowdhery, A., Castro-Ros, A., Pellat, M., Robinson, K., … Wei, J. (2024). Scaling

instruction-finetuned language models. *Journal of Machine Learning Research, 25*(1), Article 70. https://www.jmlr.org/papers/volume25/23-0870/23-0870.pdf.

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature*, *520*(7548), 429-431. https://doi.org/10.1038/520429a.

Huang, K., Mo, F.-R., Li, H.-L., Li, Y., Zhang, Y.-C., Yi, W.-J., Mao, Y.-L-., Liu, J.-C., Xu, Y.-Z., Xu, J.-N., Nie, J.-Y., & Liu, Y. (2024). A survey on large language models with multilingualism: Recent advances and new frontiers. *arXiv preprint,* arXiv:2405.10936. https://doi.org/10.48550/arXiv.2405.10936.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation*, *3*(1), 79-87. https://doi.org/10.1162/neco.1991.3.1.79.

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. https://doi.org/10.1007/s11024-019-09385-2.

Moed, H. F. (2005). *Citation analysis in research evaluation*. Springer Dordrecht. https://doi.org/10.1007/1-4020-3714-7.

Mohammadi, E., Thelwall, M., Cai, Y.-Z., Collier, T, Tahamtan, I, & Eftekhar, A. (2026). Is generative AI reshaping academic practices worldwide? A survey of adoption, benefits, and concerns. *Information Processing & Management, 63*(1), 104350. https://doi.org/10.1016/j.ipm.2025.104350.

Mukherjee, B. (2022). Evaluating the research performance of women scientists in Indian research laboratories based on Scopus citation database: A bibliometric analysis. *Malaysian Journal of Library and Information Science*, *27*(1), 57-72. https://doi.org/10.22452/mjlis.vol27no1.4.

Nowak, S., Wulff, B., Layer, Y. C., Theis, M., Isaak, A., Salam, B., Block, W., Kuetting, D., Pieper, C. C., Luetkens, J. A., Attenberger, U., & Sprinkart, A. M. (2025). Privacy-ensuring open-weights large language models are competitive with closed-weights GPT-4o in extracting chest radiography findings from free-text reports. *Radiology*, *314*(1), e240895. https://doi.org/10.1148/radiol.240895.

Qi, S.-H., Cao, Z.-Y., Rao, J., Wang, L., Xiao, J., & Wang, X.-A. (2023). What is the limitation of multimodal LLMs? A deeper look into multimodal LLMs through prompt probing. *Information Processing & Management, 60*(6), 103510. https://doi.org/10.1016/j.ipm.2023.103510.

Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q. V., Hinton, G., & Dean, J. (2017). Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint,* arXiv:1701.06538.

Sowe, S., Mou, Y.-L., Cheng, D., Kong, L.-X., Neumann, A. T., & Decker, S. (2024). Understanding open source large language models: An exploratory study. In *2024 2nd International Conference on Foundation and Large Language Models (FLLM)* (pp. 132-140). IEEE. https://doi.org/10.1109/fllm63129.2024.10852438.

Thakkar, N., Yuksekgonul, M., Silberg, J., Garg, A., Peng, N.-Y., Sha, F., Yu, R., Vondrik, C., & Zou, J. (2025). Can LLM feedback enhance review quality? A randomized study of 20k reviews at ICLR 2025. *arXiv preprint,* arXiv:2504.09737.

Thelwall, M., & Kurt, Z. (2025). Research evaluation with ChatGPT: Is it age, country, length, or field biased? *Scientometrics, arXiv preprint*, arXiv:2411.09768. https://doi.org/10.1007/s11192-025-05393-0.

Thelwall, M. (2025a). Is Google Gemini better than ChatGPT at evaluating research quality? *Journal of Data and Information Science, 10*(2), 1–5. https://doi.org/10.2478/jdis-2025-0014.

Thelwall, M. (2025b). In which fields do ChatGPT 4o scores align better than citations with research quality? *arXiv preprint*, arXiv:2504.04464. https://doi.org/10.48550/arXiv.2504.04464.

Thelwall, M. (2025c). 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.

Thelwall, M. (2025d). Research quality evaluation by AI in the era of large language models: Advantages, disadvantages, and systemic effects - An opinion paper. *Scientometrics*. https://doi.org/10.1007/s11192-025-05361-8.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). Emergent abilities of large language models. *Transactions on Machine Learning Research*, arXiv:2206.07682. https://doi.org/10.48550/arXiv.2206.07682.

Wu, M.-J., Zhang, Y., Haunschild, R., & Bornmann, L. (2025). Leveraging large language models for post-publication peer review: Potential and limitations. In S. Sargsyan, W. Glänzel, & G. Abramo (Eds.), *Proceedings of the 20th International Conference on Scientometrics & Informetrics (ISSI 2025)*, (pp. 1207-1226). International Society for Scientometrics and Informetrics. https://doi.org/10.51408/issi2025_078.

Xiao, G.-X., Lin, J., Seznec, M., Wu, H., Demouth, J., & Han, S. (2023). SmoothQuant: Accurate and efficient post-training quantization for large language models. In A. Krause, E. Brunskill, K.-H. Cho, B. Engelhardt, S. Sabato, & J. Scarlett (Eds.), *ICML' 23: Proceedings of the 40th International Conference on Machine Learning* (pp. 38087-38099). PMLR. https://proceedings.mlr.press/v202/xiao23c.html.

Zhou, R.-Y., Chen, L., & Yu, K. (2024). Is LLM a reliable reviewer? A comprehensive evaluation of LLM on automatic paper reviewing tasks. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N.-W. Xue (Eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 9340-9351). ELRA and ICCL. https://aclanthology.org/2024.lrec-main.816/.

**Appendix 1: Sample report (emphasis and redactions added)**

**## Assessment of: "[redacted]"**

****Overall Score: 4*****
****Justification:**** This article represents world-leading research in its field, demonstrating exceptional quality across all three assessed criteria: originality, significance, and rigour. The discovery and comprehensive characterization of [redacted] is a landmark achievement with substantial implications for evolutionary biology and [redacted] research.

****1. Originality (4*):****
The discovery of [redacted] is a highly original contribution. While the search for such a species has been ongoing, its successful identification and detailed characterization represent a significant breakthrough. The research doesn't simply present new data; it *creates* a new system for investigation. The work goes beyond mere description, delving into genomic analysis, identifying unique features like the expansion of transposable elements and massive gene losses in [redacted]. This isn't just a descriptive genomic report; it's a comparative genomic analysis immediately framing evolutionary questions. The development of genetic and molecular techniques specifically for [redacted] further underscores the innovative nature of this work, providing the tools necessary for future research. The use of a novel ecological niche [redacted] as the source of the new species also demonstrates a creative approach to discovery.

****2. Significance (4*):****
The significance of this discovery is substantial. [redacted] is a cornerstone of biological research, and the availability of [redacted] opens up unprecedented opportunities for comparative evolutionary studies. The authors explicitly state this potential, and it is readily apparent. The ability to compare genomes and phenotypes will allow researchers to dissect the genetic basis of adaptation, identify conserved and divergent pathways, and gain deeper insights into the evolution of development, behaviour, and genome structure. The findings regarding transposable element expansion and chemoreceptor gene loss are particularly significant, offering potential explanations for species divergence and adaptation to specific environments. The work has the potential to influence a broad range of research areas, from molecular biology and genetics to ecology and evolutionary biology. The creation of a new model system will undoubtedly stimulate further research and generate new knowledge.

****3. Rigour (4*):****
The research demonstrates exceptional intellectual coherence and methodological rigour. The study employs a multi-faceted approach, integrating morphological, developmental, behavioural, and genomic analyses. The genome sequencing and assembly [redacted] are presented with sufficient detail to assess their quality. The identification of unique genomic features is supported by robust bioinformatics analyses. The authors clearly articulate the methods used and the rationale behind their experimental design. The comparison between [redacted] and [redacted] is systematic and well-documented. The development of genetic and molecular tools for the new species indicates a commitment to ensuring the reproducibility and accessibility of the research. The logical flow of the paper, from discovery to characterization to potential applications, is highly coherent and demonstrates a strong understanding of the field.

****Concluding Remarks:****

This article is a truly outstanding piece of research. It represents a significant advance in our understanding of nematode evolution and provides a valuable new resource for the scientific community. The combination of originality, significance, and rigour firmly places this work at the forefront of its field, deserving of the highest possible rating.