Is Google Gemini better than ChatGPT at evaluating research quality?

Mike Thelwall

School of Information, Journalism and Communication, University of Sheffield, Sheffield, UK

Citation: Thelwall, M. (2025). 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

DOI: 10.2478/jdis-2025-0014 **CSTR:** 32295.14.jdis-2025-0014

Received: Dec. 9, 2024; Revised: Dec. 24, 2024; Accepted: Dec. 25, 2024

ABSTRACT

Google Gemini 1.5 Flash scores were compared with ChatGPT 4o-mini on evaluations of (a) 51 of the author's journal articles and (b) up to 200 articles in each of 34 field-based Units of Assessment (UoAs) from the UK Research Excellence Framework (REF) 2021. From (a), the results suggest that Gemini 1.5 Flash, unlike ChatGPT 4o-mini, may work better when fed with a PDF or article full text, rather than just the title and abstract. From (b), Gemini 1.5 Flash seems to be marginally less able to predict an article's research quality (using a departmental quality proxy indicator) than ChatGPT 4o-mini, although the differences are small, and both have similar disciplinary variations in this ability. Averaging multiple runs of Gemini 1.5 Flash improves the scores.

1 Introduction

This article compares Gemini 1.5 Flash with ChatGPT 4o-mini for the task of estimating journal article research quality on 51 information science articles with author-assigned quality scores (Thelwall, 2025) and 6,648 from all fields (Thelwall & Yaghi, 2024). The Flash/mini versions of these are the cut down reduced cost variants that seem to have similar performance to the full models but are cheaper and hence more practical for large scale experiments and applications.

- RQ1: Which inputs (title, title and abstract, full text without references, full text, PDF) give the best results for Google Gemini 1.5 Flash for research quality evaluation?
- RQ2: Does averaging multiple repetitions of queries improve the value of Google Gemini

1.5 Flash academic score predictions?
- RQ3: Are there disciplinary differences in the usefulness of Google Gemini 1.5 Flash?

2 Methods

2.1 Data sets

The first data used was a set of 51 published and unpublished journal articles written by the author, as used in previous studies (Thelwall, 2025). Each article has an associated quality score from the author, using the Research Excellence Framework (REF) scale of 1* (recognised nationally), 2* (recognised internationally), 3* (internationally excellent) and 4* (world leading) for originality, significance, and rigour. Articles were given fractional scores (1.5*, 2.5* or 3.5*) when they fell between two thresholds. These scores are private.

The data used for the last two research questions was the collection of up to 200 journal articles from the highest and lowest scoring departments in each of the 34 UoAs in REF2021, as used in a previous study for ChatGPT (Thelwall & Yaghi, 2024). Although each article is individually scored in the REF by two experts, with scores norm-referenced within and between UoAs, only aggregate departmental scores are published. Each article was therefore associated with the departmental mean REF score for its UoA. Assuming that there are only weak departmental biases in the system, this substitution of actual article quality scores with a departmental proxy would have a damping effect on any correlations with potential quality indicators, reducing their magnitude but not changing their sign.

2.2 Gemini prompts and scoring

Each article was uploaded to Gemini 1.5 through its API. The main versions at the time of the experiment (October 2024) were: 1.5 Pro (gemini-1.5-pro-002), the "best performing" multimodal model; 1.5 Flash (gemini-1.5-flash-002) a cheaper variant "balancing performance and cost"; and 1.5 Flash-8B (gemini-1.5-flash-8b-001), the "fastest and most cost-efficient" (ai.google.dev/gemini-api/docs/models/gemini). Gemini 1.5 Flash was chosen because the more expensive alternative may not be practical for research evaluations. It is equivalent to ChatGPT 4o-mini, the cut down version of ChatGPT 4o used in most previous similar studies.

The system instructions used were as previously used for ChatGPT (Thelwall, 2025). They are the REF2021 scoring guidelines for assessors in Main Panel D (arts and humanities but including information science) for the first dataset and the appropriate guidelines for the second dataset (UoAs 1-6: Main Panel A, health and life sciences; UoAs 7-12: Main Panel B, physical sciences and engineering; UoAs 13-24: Main Panel C, social sciences; UoAs 25-34: Main Panel D, mainly arts and humanities). Article scores were extracted from Gemini reports with an extended version of the software designed to extract scores from ChatGPT reports for the same task (Webometric Analyst, AI menu; ChatGPT: Extract scores from reports; github.com/MikeThelwall/Webometric_Analyst).

2.3 Analysis

As in the previous studies, the main statistical test was correlation between the Gemini scores and the article quality scores (first dataset) or departmental quality proxy for article quality (second dataset). Correlation is the most appropriate test because the results can easily be scaled (Thelwall, 2025) if Gemini tends to score high or low. Spearman correlations were used since the data are fundamentally ranks and for comparability with previous papers.

3 Results

3.1 RQ1, RQ2: Comparison between inputs and numbers of repetitions

The maximum Spearman correlation with author scores for Gemini 1.5 Flash is 0.645 for PDF and 0.549 for truncated text (no references or tables), after 30 iterations (Figure 1). The correlations are lower than previously found for ChatGPT in all cases except Full text which was not checked in ChatGPT, and PDF, for which the ChatGPT 4 correlation was 0.509 after 15 iterations (Thelwall, 2025) (the Gemini 1.5 Flash correlation was 0.613 after 13 iterations). Thus, Google Gemini may be relatively better than ChatGPT at analyzing long documents, including full text. The differences are not statistically significant, however. For example, a bootstrapped 95% confidence interval for the PDF Spearman correlation is (0.38, 0.81). This should be interpreted as the likely range of values for the correlation for a similar but different set of articles. Comparing ChatGPT 4o-mini with Gemini 1.5 Flash, the optimal strategy found so far is using Gemini 1.5 Flash with PDFs, at least in terms of the highest correlation with article quality scores. Of course, this is only a tentative conclusion, given the small sample size and small correlation differences involved.

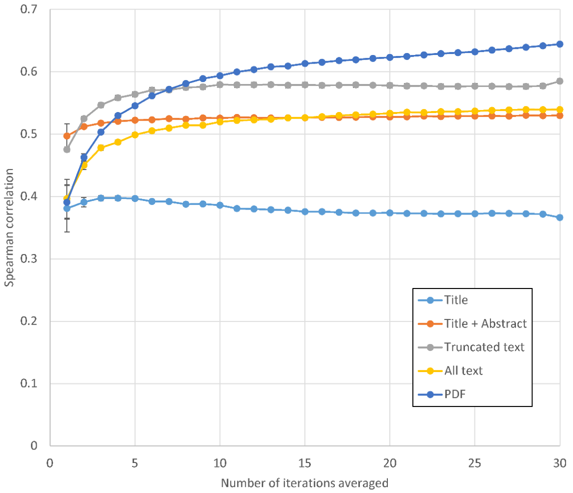

Figure 1. Spearman correlations between Gemini 1.5 Flash scores and the author's scores for 51 library and information science articles, against the number of repetitions averaged. Each line represents a different amount of input. Error bars are 95% confidence intervals for averaging within the data collected.

3.2 RQ3: Comparison between disciplines

For the REF data, Gemini 1.5 Flash correlations with departmental mean REF2021 scores tend to be similar in magnitude to the correlations for ChatGPT 40-mini. Nevertheless, the overall mean of the correlations is marginally higher for ChatGPT 4o-mini (rho=0.409) than for Gemini 1.5 Flash (rho=0.399) (Figure 2). In both cases, the correlations are relatively weak for most arts and humanities but strong in all health, life and physical sciences and some social sciences. Clinical Medicine has an anomalously low Gemini correlation, but it is positive, unlike the corresponding ChatGPT correlation.

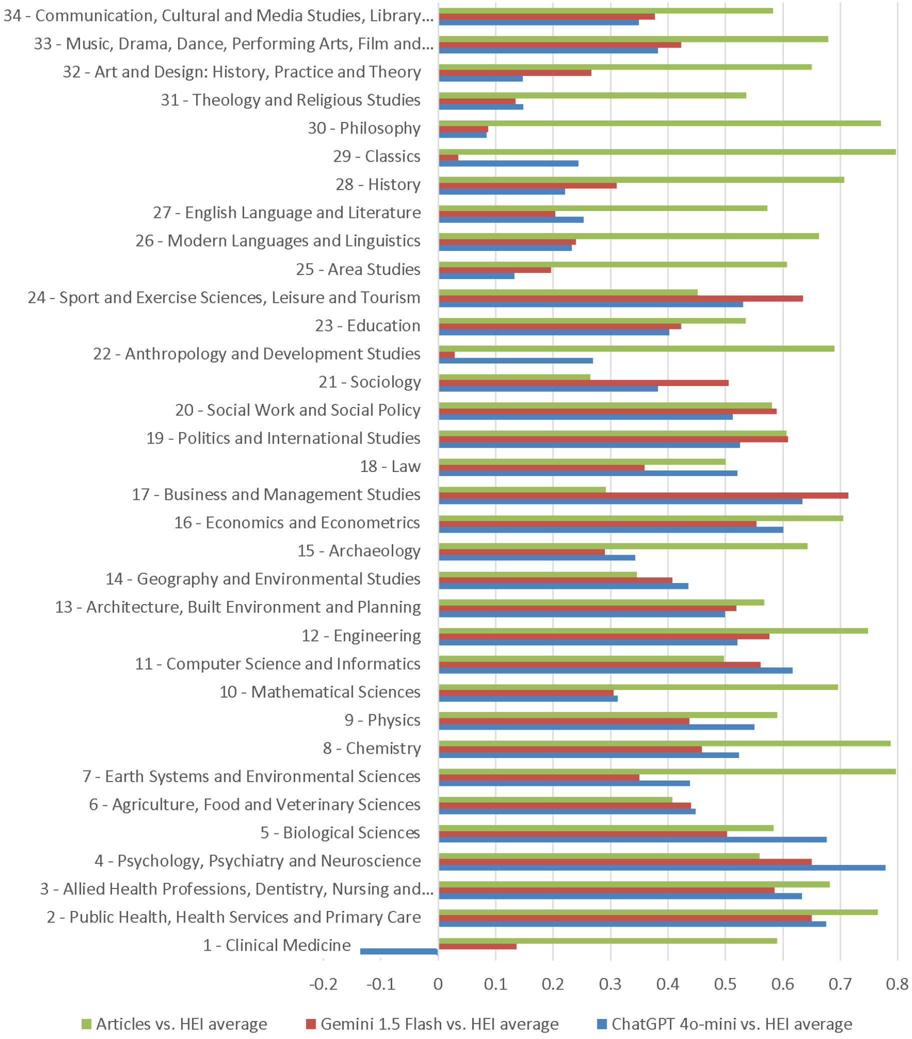

Figure 2. Spearman correlations between Gemini 1.5 Flash scores and departmental average REF2021 scores. Also shown are equivalent correlations from ChatGPT 4o-mini and, as a benchmark, the correlation between article scores and departmental average REF2021 scores. Error bars are 95% confidence intervals for the assumed infinite population of similar articles.

References

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

Thelwall, M., & Yaghi, A. (2024). In which fields can ChatGPT detect journal article quality? An evaluation of REF2021 results. https://arxiv.org/abs/2409.16695