Can ChatGPT replace citations for quality evaluation of academic articles and journals? Empirical evidence from library and information science

Mike Thelwall, University of Sheffield, UK

Purpose: Whilst citation-based indictors have been recommended by librarians to support research quality evaluation, they have many acknowledged limitations. ChatGPT scores have been proposed as an alternative, but their value needs to be assessed.

Design/methodology/approach: Mean normalised ChatGPT scores and citation rates were correlated for articles published 2016-2020 in 24 medium and large Library and Information Science (LIS) journals on the basis that positive values would tend to support the usefulness of both as research quality indicators. Word association thematic analysis was employed to compare high and low scoring articles for both indicators.

Findings: There was a moderately strong article-level Spearman correlation of rho=0.448 (n=5925) between the two indicators. Moreover, there was a very strong journal-level positive correlation rho=0.843 (n=24) between the two indicators, although three journals had plausible reasons for being relatively little cited compared to their ChatGPT scores. ChatGPT seemed to consider research involving libraries, students, and surveys to be lower quality and research involving theory, statistics, experiments and algorithms to be higher quality, on average. Technology adoption research attracted many citations but low ChatGPT scores, and research mentioning novelty and research context was scored highly by ChatGPT but not extensively cited.

Originality: This is the first evidence that ChatGPT gives plausible quality rankings to library and information science articles, despite giving a slightly different perspective on the discipline.

Practical implications: Academic librarians should be aware of this new type of indicator and be prepared to advise researchers about it.

Keywords: ChatGPT, Large Language Models; LLMs; Research Evaluation; Scientometrics; Bibliometrics

Introduction

Although Library and Information Science (LIS) includes the specialities of bibliometrics and research evaluation, evaluating the quality of published academic research and journals is still time consuming, difficult, and controversial for LIS academics and practitioners. For this reason, citation-based indicators like Journal Impact Factors (JIFs), article citation counts, or career citations may sometimes be recommended by academic librarians to support research quality evaluators, especially if they lack the expertise or time for a rigorous evaluation. This is important for information schools, which naturally contain a range of different types of expertise, as is evident from the staff profiles in their websites. They may also seek staff in new areas to fill teaching gaps (e.g., to support new data science courses: Urs & Minhaj, 2023), so may lack the expertise to evaluate the research of some of their applicants.

This article evaluates ChatGPT quality scores as a source of evidence about the quality of LIS journals (i.e., average quality of the articles published in them) and journal articles. Whilst there have been many previous attempts to evaluate research quality with citation analysis or traditional artificial intelligence approaches (e.g., Wang et al., 2024), the recent

successes at various tasks of ChatGPT, including research evaluation (Thelwall, 2024ab), make it a particularly promising candidate.

The article has two overlapping and partly conflicting motivations. A first goal is to assess whether ChatGPT scores for LIS research are meaningful rather than random in the sense of positively correlating with citation-based indicators (whatever their limitations) at the article and journal level. This is the motivation behind the first research question.

- RQ1: Does ChatGPT give higher scores to more cited articles and to articles in journals in library and information science journals that publish more cited work?

A second goal is to use the perspective of ChatGPT scores to gain insights into the limitations of citation-based indicators within LIS and, conversely, to use citation-based indicators to gain insights into the limitations of ChatGPT scores. This may also give a new perspective about the nature of high-quality LIS research. The second question addresses this goal.

- RQ2: Which types of library and information science journal articles tend to get high or low scores from ChatGPT and/or many or few citations?

Background

Citation analysis

Citation counts, and indicators based on citation counts, are widely used to support research evaluations and even sometimes as the sole evidence for some types of evaluations. Their primary advantages compared to human expert judgement are speed (it is faster to calculate numbers than to read an article), objectivity (the numbers are facts, not opinions) and fine-grained nature (citation counts can vary widely, whereas reviewers usually score on a very restricted scale). Common applications include ranking or evaluating journals based on their citation rates (e.g., JIFs), ranking or evaluating scholars based on their career citations or h-index (a sexist and ageist practice), and using an article's citation count as a quick indicator of its value, perhaps during a literature search (de Bellis, 2009; Moed, 2006).

The theory behind the use of citation-based indicators is that citations serve to acknowledge prior works that have influenced a new study (Merton, 1973). On this basis, the more citations an article or journal has, the more influential it has been. This hypothesis is problematic because citations are routinely used for many other purposes including criticism and background context, the selection of citations is biased by many factors, and perhaps most fundamentally, citations do not reflect non-scholarly impacts (MacRoberts & MacRoberts, 2018). In response to these criticisms, it has been argued that, in the absence of biasing factors, the irrelevant citation reasons may tend to cancel each other out when citations are aggregated over large sets of texts (van Raan, 1998).

In support of the above claim that biasing factors and irrelevant citations may tend to aggregate out on a larger scale, appropriately field and year normalised citation rates correlate positively with expert quality ratings in almost all fields, albeit with varying strengths (Thelwall et al., 2023c). This gives evidence that citation-based indicators can convey genuine, if sometimes weak, evidence of research quality. Of course, research evaluators may be influenced by an article's citation count or the citation impact of the publishing journal, undermining the validity of this evidence. Nevertheless, the positive correlation suggests that citation-based indictors can theoretically be used to support expert judgement in some ways, such as guiding experts for papers that they do not understand, to cross-check judgements, to look for human error or bias, or to resolve disagreements between expert reviewers.

Despite the need to support and carry out research evaluations and a role in developing, teaching, and supporting bibliometrics, the LIS field has been at the forefront of warning against the inappropriate use of bibliometric indicators because of the limitations mentioned above. Cautionary advice has been expressed in the Metric Tide report (Wilsdon et al., 2015), the Leiden Manifesto (Hicks et al., 2015) and by CoARA (coara.eu), as well as by supporting initiatives like DORA (sfdora.org).

Within the speciality of bibliometrics, there are variations of opinions about the fundamental value of citation-based indicators (Rushforth & Hammarfelt, 2023). One of the main recognised limitations of citation-based evaluations is that research quality is usually believed to encompass originality, rigour, scholarly impact, and societal value (Langfeldt et al., 2020), whereas citations primarily reflect scholarly impact (Aksnes et al., 2019). This means that any evaluation using citation-based indicators risks ignoring three important components of research quality: originality, rigour, societal value.

Another problem is that citations take several years to mature enough to provide useful evidence (Wang, 2013) whereas recent research is often the most salient in evaluations. For example, a citation-based research evaluation conducted in 2025 might only be able to evaluate research published before 2023.

Artificial Intelligence (AI) and Large Language Models (LLMs) for research evaluation

Artificial intelligence methods have occasionally been used as an alternative to citation analysis for automated quality evaluations, although there have been many studies that have evaluated the potential of AI. Before LLMs, AI methods typically harnessed citation data from the article analysed, and sometimes also from its author(s), and publishing journal, as well as exploiting metadata like title and abstract text and reference lists (Kousha & Thelwall, 2024). The most powerful inputs for AI systems were article and journal citation rates, however (Thelwall et al., 2023d), so pre-LLM AI methods have essentially been advanced types of citation analysis algorithm, with the same limitations as citations.

In theory, Large Language Models like ChatGPT can address both main limitations of citation-based indicators (i.e., that they ignore originality, rigour and societal impact, and that they are not applicable for studies published in the most recent three years). They can be fed system instructions that explicitly tell them to evaluate originality, rigour, scholarly impact, and societal significance and can be asked to evaluate research of any age. Moreover, ChatGPT can provide useful advice on journal article submissions (Liang et al., 2024b; Zhou et al., 2024) and it is already sometimes used by reviewers for language polishing, if not insights into the reviewed articles (Liang et al., 2024a).

Although the above paragraph is mostly theoretical, empirical evidence now suggests that ChatGPT can produce useful quality evaluations of published academic journal articles across all of academia, although the strength of association (correlation) between (an indicator of) expert scores and ChatGPT scores varies between fields (Thelwall & Yaghi, 2024; Thelwall et al., 2025). In most fields, ChatGPT scores seem to outperform bibliometrics and traditional machine learning as a research quality indicator (Thelwall & Yaghi, 2024), perhaps because LLMs can be asked to consider wider dimensions of research quality. Of course, LLMs are also imperfect, and research is needed to evaluate them for narrower fields than the 34 broad groupings previously assessed, including LIS, as well as to gain insights into the types of research that LLMs value.

Methods

The research design was to obtain a set of medium or large library and information science journals (operationalised heuristically as at least 250 articles over five years) for a recent period and then a) compare their average citation rates with their ChatGPT scores and b) identify topics within the articles published in these journals that are associated with high and low citation rates and/or ChatGPT scores. Medium or large journals were needed to give a reasonable degree of precision to the statistical analyses. A period of five years is small enough to avoid large systemic variations due to changes in the field, whilst increasing the statistical power of the tests by including multiple years.

The choice of the number of years to examine in a bibliometric study is a necessarily heuristic decision, but topics within the field tend to be relatively stable over time, although individual issues can appear within single years, changing part of the field (Figuerola et al., 2017). Previous studies have also considered five years to be a reasonable period for a bibliometric analysis, even going so far as to split longer periods into five-year chunks for analysis (Siddique et al., 2021). Although some research designs use power analysis to decide sample sizes, this study is constrained by the amount of data available, and the sample choice necessarily needs to take this into account.

Dataset

The period 2016-2020 was chosen for the sample, giving almost four complete years (from data collection in November 2024) to accrue citations and substantially longer for most articles. This is sufficient for citation counts to mature in most fields (Wang, 2013). Heuristically, a five-year span 2016-2020 includes enough articles that medium sized journals can be included with enough articles. The records for the documents of type journal article in the Scopus narrow field Library and Information Sciences were then downloaded using Scopus's Application Programming Interface (API) during November 3, 2024.

Some of the article records had no abstracts or had very short abstracts. In many cases this seemed to be due to the documents being news stories or other short form contributions rather than standard journal articles. To avoid making meaningless comparisons between journals based on Scopus classification anomalies, all records for documents with descriptions in the shortest 10% (less than 644 characters) were excluded. The 10% threshold seemed adequate do distinguish between mainly short form contributions and mainly long form contributions, although this was judged subjectively.

A second subjective filtering step was also used. Since Scopus usually assigns multiple narrow fields to journals, the LIS category includes journals that have little connection to the main field. The journal list was therefore pruned of titles seemed not to be part of mainstream LIS because they primarily focus on a different field, such as computing or health, consulting journal websites and lists of published articles to help with the decision. The omitted journals (that would have met the size threshold below) are, *Big Data and Society; Data Analysis and Knowledge Discovery; Development and Learning in Organizations; Education and Information Technologies; IEEE Transactions on Information Theory; Information Technology and People; International Journal of Data Mining and Bioinformatics; International Journal of Geographical Information Science; Journal of Chemical Information and Modeling; Journal of Cheminformatics; Journal of Health Communication; Journal of Information and Computational Science; Journal of Information Science and Engineering; Language Resources and Evaluation; Notes and Queries; Personal and Ubiquitous Computing; Scientific Data; Social*

Science Computer Review. Including journals that primarily focused on a different field would undermine attempts to find patterns within LIS.

The number of articles per journal was then counted, with an original target of retaining all journals with at least 250 articles, to enable robust statistical analyses. This was relaxed to allow five smaller journals to give a more balanced set (*DESIDOC Journal of Library and Information Technology*: 248 articles; *Evidence Based Library and Information Practice*: 247 articles; *Journal of Library Administration*: 231 articles; *Journal of the Medical Library Association*: 231 articles; *Serials Librarian*: 218 articles). For journals with more than 250 articles, 250 articles were selected at random using a random number generator to avoid an unbalanced set for RQ3. The set still has some bias since smaller humanities-oriented journals may have been disproportionately excluded and five of the selected journals had fewer than 250 articles.

ChatGPT scores

The most useful score predictions from the ChatGPT API so far have been obtained by priming it with system instructions that mirror Research Excellence Framework (REF) 2021 guidelines (see Appendix), then submitting a user prompt of the form "Score the following article:", followed by the article title, then "Abstract" on a new line (by adding "\n" to the submitted Json) and the article abstract, also on a new line (Thelwall, 2024ab). This prompt works by defining research quality in a moderately discipline sensitive way (the prompt covers the arts and humanities and some social sciences) and then giving guidelines for the four quality criteria used. ChatGPT 3.5+ is optimised for natural language prompts so the use of guidelines for human reviewers is a logical approach. Submitting full texts does not seem to improve the results, at least in terms of correlations with an external measure of quality, and the variations of this prompt tested so far have not improved it, with parameter variations making it worse (Thelwall, 2024ab). Submitting the same query multiple times and averaging the results tends to improve the score, however (Thelwall, 2024ab).

Since the optimal ChatGPT strategy does not involve the full text of an article, only its title and abstract, ChatGPT cannot be claimed to be "measuring" or "evaluating" the quality of articles when it scores them. Instead it is *guessing* their quality, presumably from the extent to which the title/abstract descriptions match the REF2021 quality guidelines, taking into account ChatGPT's wider knowledge of research issues from its general reading. The empirical evidence cited above shows that these guesses correlate positively with expert judgements in almost all fields, and correlate more strongly than citation-based indicators. It is therefore reasonable to use ChatGPT scores as research quality *indicators*, whilst acknowledging that they are not research quality *measures* (the same is true for citations). Since expert judgement is usually accepted as the gold standard for research quality (although experts frequently disagree), the value a research quality indicator lies primarily in the extent to which it correlates positively with human judgement. It is also important to consider whether it incorporates systematic biases (e.g., against qualitative research), and whether it is gameable.

The strategy above (ChatGPT scores from article titles and abstracts based on REF2021 guidelines) was used for the journal article sample, with the appropriate REF guidelines (i.e., from Main Panel D, which includes LIS) and submitting each query five times. Although more repetitions would improve the value of the predictions, each additional one adds relatively little and five seems adequate. The queries were submitted through the API 1-3 November 2024 to ChatGPT 4o-mini. This is a smaller version of the main general ChatGPT model at the time, 4o, but has similar performance and is more practical because it is much cheaper

(Thelwall, 2024b). ChatGPT does not train on API data (OpenAI, 2024ab) and UK law allows lawfully accessed journal article abstracts to be used for research purposes (Kelly, 2024), so the procedure was legal.

Since ChatGPT scores tend to be higher for more recent years (Thelwall & Kurt, 2024), when combining multiple datasets the publication year should be corrected for. Thus, the average of all ChatGPT scores was calculated separately by year and, for years other than the middle year 2018, an offset was calculated: the 2018 ChatGPT mean minus year ChatGPT mean. This was then added to all ChatGPT articles from that year. After this, the "offset" ChatGPT mean was the same for each year.

Journal mean ChatGPT score: This is the mean of the offset ChatGPT scores for articles in the journal.

Citation rates

Citation datasets are typically skewed, with citation counts also varying by year. A strategy was therefore needed for both issues. Skewing (a few high citation counts) was dealt with by log normalisation with the formula ln(1+citations), giving a Log Citation Score (LCS) for each article. Here, 1 is added because ln(0) is undefined and there are some uncited articles. To normalise by year, the mean LCS for each year was calculated and each LCS divided by the appropriate year mean, giving the Normalised LCS (NLCS). By design, this is 1 for an article with a log citation count equal to the mean for its year. Values above 1 indicate an article being more cited than average and values less than 1 indicate less cited than average. It is fair to compare NLCS between years or to average them across years since they are time normalised (Thelwall, 2017).

Journal mean citation rate: This is the Mean NLCS (MNLCS) for articles in the journal. By design it is skew corrected (so a more precise central tendency estimate) and time corrected.

RQ1: ChatGPT scores against citation rates for articles and journals

Offset ChatGPT scores were correlated against NLCS (normalised citations) to assess the extent to which they agree. Journal mean citation rates were also correlated with journal mean ChatGPT scores to identify whether more cited journals tended to receive higher ChatGPT scores. Spearman correlations are the primary value since the article-level data failed Shapiro-Wilk normality tests for at least one variable in all cases, but Pearson correlations are reported as descriptive statistics. Confidence intervals were calculated by bootstrapping.

RQ2a: Journal articles tending to get high or low scores from ChatGPT

Although there are many methods to find patterns in document collections, Word Association Thematic Analysis (WATA) (Thelwall, 2023) was chosen because its focus is on finding differences between two sets of texts, it is statistically based to reduce the chance of finding spurious patterns, and it has high face validity because it finds themes by manually analysing terms in their contexts. Other well-known methods, such as clustering and topic modelling, do not identify statistically significant themes in the differences between two sets of texts.

Sets of terms associating with high ChatGPT scores: For the first WATA, the article titles and abstracts were fed into the text analytics software Mozdeh (github.com/MikeThelwall/Mozdeh) along with their ChatGPT adjusted scores (x1000 and rounded, since Mozdeh only accepts whole numbers). Mozdeh was then used to list all the

words occurring disproportionately often in the (titles and abstracts of) the articles with above median (2.86) ChatGPT scores compared to the rest. The degree of disproportionality was assessed with a 2x2 chi square test in Mozdeh, and with Benjamini-Hochberg familywise error correction applied to reduce the chance of spurious false positives from multiple simultaneous tests (Benjamini & Hochberg, 1995).

Term contexts: The words identified above as occurring disproportionately often in high scoring articles were examined by reading a random sample of 20 articles (title and abstract) containing them and then assigning the most common context within the corpus. This was supported by secondary word association analyses: identifying words that tended to co-occur with them disproportionately often (again using Mozdeh). This is important not only because some words are polysemic (e.g., "staff" could mean employees or a stick) but also because terms were often used within a particular context (e.g., "university" was usually used in the context of university libraries; "electronic" usually referred to resources provided by academic libraries; "India" was usually used to describe the location of a study; "was" was used for many purposes but reflected abstracts written in the past tense). This was achieved for all terms that occurred statistically significantly more in the higher scoring half of the dataset with $p<0.001$.

Theme identification: The terms were iteratively and reflexively clustered into themes of similar contexts. This involved merging similar term contexts to examine whether they were fundamentally different or were parts of a more general concept. For example, this stage produced a theme that was a cluster of terms relating to academic resource provision from university libraries, and a second theme that was a cluster of terms relating to university student education. Although these two themes have some overlap, they were sufficiently different to be distinguished. The main result of this stage was a set of themes associated with articles that tended to have above average ChatGPT scores in the dataset.

The above process was then repeated for the articles with below average ChatGPT scores.

RQ2b: Which types of library and information science journal articles tend to attract many or few citations?

The RQ2 methods were repeated for articles that were cited above or below the median NLCS. To ensure the maximum consistency between schemes, the high ChatGPT and high citation rate WATAs were conducted in parallel using the same categories whenever possible. Similarly, the low ChatGPT and low citation rate WATAs were also conducted together. Themes associated with journal heading styles are not reported.

Results

The journal level scores are summarised in Table 1.

Table 1. Descriptive statistics for the journals analysed. Source: Author's own work.

Journal	Mean GPT	Mean NLCS	Mean abstract length	Articles
DESIDOC Journal of Library and Information Technology	2.128	0.679	1136	248
Electronic Library	2.582	0.931	1854	250
Evidence Based Library and Information Practice	2.616	0.424	2613	247
Government Information Quarterly	2.937	1.557	1264	250
Information Communication and Society	2.988	1.330	1294	250
Information Development	2.528	1.016	1182	250
Information Processing and Management	2.995	1.462	1473	250
Information Systems Research	3.107	1.549	1548	250
International Journal of Information Management	2.869	1.777	1233	250
Journal of Academic Librarianship	2.512	0.976	1131	250
Journal of Documentation	2.915	1.010	1724	250
Journal of Information Science	2.864	1.147	1253	250
Journal of Informetrics	2.981	1.312	1224	250
Journal of Librarianship and Information Science	2.477	0.976	1211	250
Journal of Library Administration	2.305	0.709	846	231
Journal of Modern Information	2.284	0.286	1143	250
J.Association for Information Science & Technology	2.966	1.167	1206	250
Journal of the Medical Library Association	2.427	0.918	1486	231
Library Philosophy and Practice	2.061	0.347	1390	250
Online Information Review	2.833	1.173	1744	250
Proc. Association for Information Science & Technology	2.661	0.569	1072	250
Profesional de la Informacion	2.518	0.956	1058	250
Scientometrics	2.803	1.144	1321	250
Serials Librarian	2.259	0.481	928	218
All	**2.655**	**1.000**	**1350**	**5925**

RQ1: ChatGPT scores against citation rates for articles and journals

There was a very strong positive Spearman correlation between normalised journal citation rates (MNLCS) and mean offset ChatGPT scores of rho=0.843 (Table 2). This might reflect aggregation effects from a moderate positive correlation at the article level of rho=0.448. Thus, there is very strong evidence to support a positive answer to both parts of RQ1, although this is not a cause-and-effect claim.

Abstract length may be important as an input for ChatGPT. It may also affect citation rates if short form articles were included in the dataset and these tended to have shorter abstracts. At the article level the correlation between abstract length and ChatGPT scores is rho=0.245 and at the journal level the correlation is rho=0.448. In contrast, citation rates corelate more weakly with abstract length (rho=0.116 for articles, rho=0.276 for journals) so there may be a small stylistic advantage with ChatGPT for articles and journals with longer abstracts.

Table 2. Article-level and journal level correlations for 24 Scopus-indexed LIS journals and their articles 2016-2020. Source: Author's own work.

Variables correlated	**Spearman**	**Pearson***	**Sample**
Journal MNLCS vs ChatGPTn	0.843 (0.639, 0.942)	0.813	24
Journal MNLCS vs abstract length	0.276 (-0.121, 0.684)	-0.024	24
Journal ChatGPTn vs abstract length	0.423 (0.006, 0.709)	0.246	24
Article NLCS vs ChatGPTn	0.448 (0.424, 0.467)	0.442	5925
Article NLCS vs abstract length	0.116 (0.091, 0.141)	-0.007	5925
Article ChatGPTn vs abstract length	0.245 (0.221, 0.270)	0.206	5925

*Included as descriptive statistics but at least one variable is not normally distributed.

Nearly all articles have ChatGPTn scores in the approximate range 2* to 3.3*, with 3* being the default (Figure 1). Thus, ChatGPT seems reluctant to give the highest and lowest available scores to LIS articles. One article received a 1* from ChatGPT (before correction) and this seemed to be a personal opinion piece that should not have been classified as an academic journal article. The next two lowest scoring articles were workshop reports but the fourth lowest scoring article was genuine research. The highest scoring article, “Folding and unfolding: Balancing openness and transparency in open source communities” did seem to be excellent from its abstract. Overall, however, the ChatGPT results need scaling to transform them to the usual REF range.

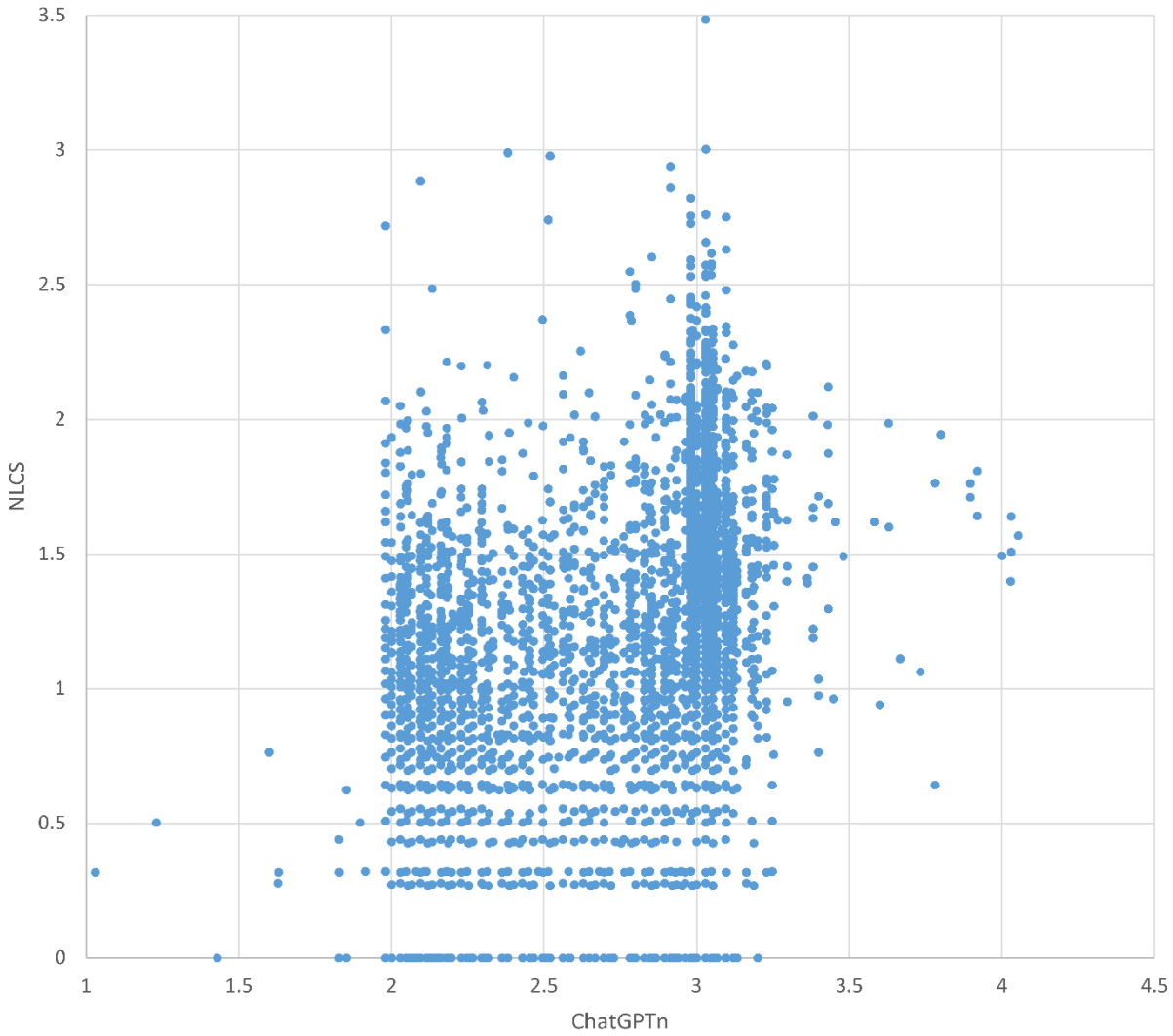

Figure 1. Article citation rates against normalised ChatGPT scores for 24 medium and large LIS journals 2016-2020, for articles without short abstracts (at least 644 characters). ChatGPTn values can be above 4* after the year normalisation. Source: Author's own work.

At the journal level, the correlation between average citation rates and ChatGPT scores reflects broad agreement between the two and a roughly linear trend, although some journals are outliers (Figure 2).

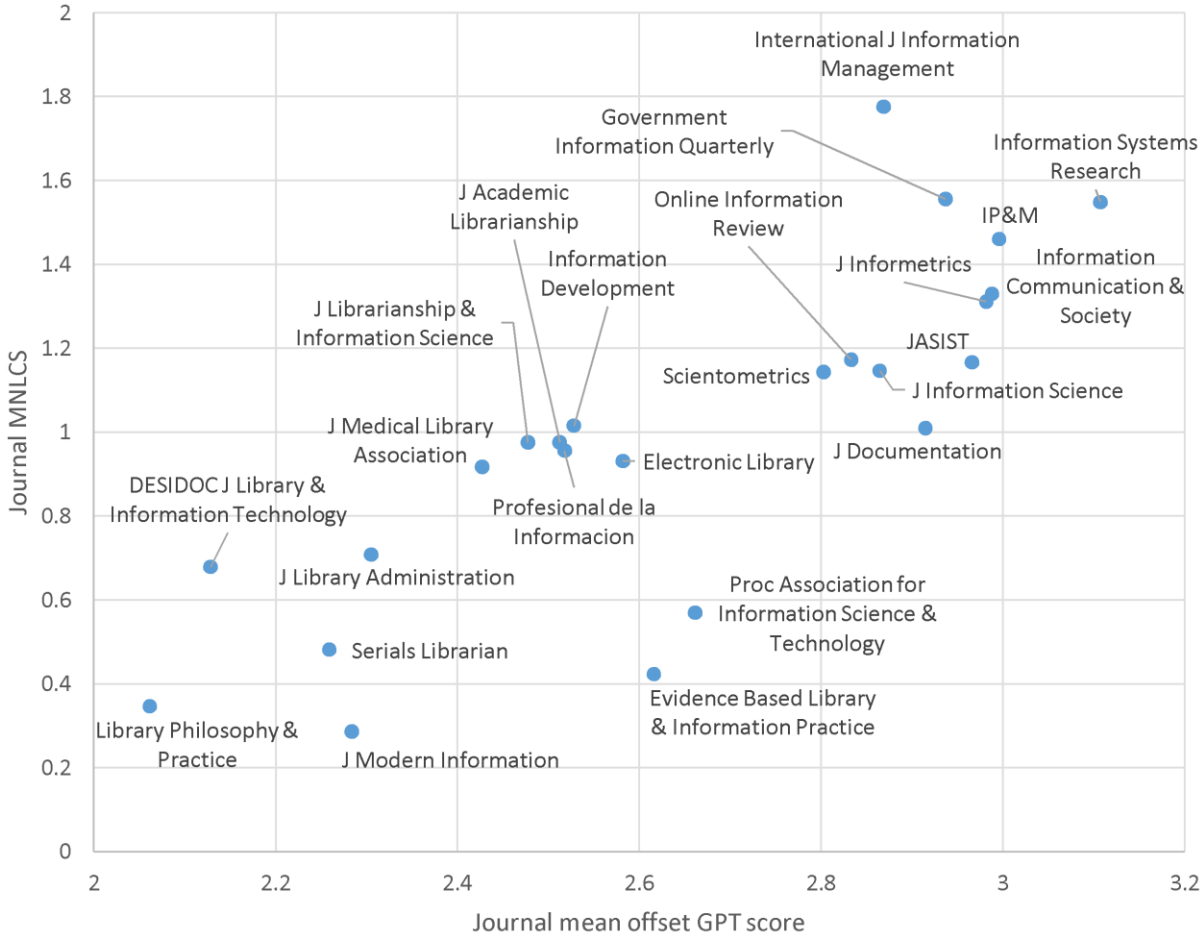

Figure 2. Journal citation rates against journal ChatGPT scores for 24 medium and large LIS journals 2016-2020, for articles without short abstracts (at least 644 characters). Source: Author's own work.

RQ2: Journal articles tending to get high or low ChatGPT scores or citation rates

There was a substantial overlap in the types of articles that tended to get high ChatGPT scores or citation rates, with 9 of the 12 WATA themes identified matching both (Table 3). In general, work mentioning theory, empirical evidence and with complex arguments tended to score well and the high scoring topics included business, government, and social media/networks.

Despite the above commonalities, ChatGPT seemed to particularly value mentions of novelty or the context of research in a way that citation rates did not. This makes sense in that ChatGPT was explicitly told to consider originality and significance (as well as rigour). In contrast, technology adoption issues tended to be more cited without achieving higher ChatGPT scores. It is possible that research into this area tends to address emerging issues (because adoption is important) and therefore might get a citation advantage for early publishing on a topic, with more mature research on the same topic tending to cite it.

Table 3. Themes found from WATA for title and abstract terms associated with high ChatGPT scores or high citation rates for LIS articles from 24 journals 2016-2020. Source: Author's own work.

Theme	**Statistically significant terms**	**ChatGPT**	**NLCS***
Novel contribution	novel	***	
Wider context of study	Context	***	
Theory	Model, proposed, approach, theory, framework, empirical	***	(***)
Experiments and algorithms	Experiment, algorithm, dataset, performance	***	(***)
Understanding or explanation	How, can, on	***	(***)
First person style	We, our, propose, show, that	***	***
Statistics	Effect, influence, affect, antecedent, structural	***	***
Complex argument	However, when, suggest	***	***
Social media/networks	Social, Twitter, media, network	(***)	***
Business	Firm, customer, consumer,	(***)	***
Government, democracy	Citizen, e-government	(***)	***
Technology adoption	Adoption		***

*Three stars indicates at least one term statistically significant with $p<0.001$. No stars indicates that no term is statistically significant with $p<0.05$. Brackets () indicate that the theme appears subjectively to be less prominent than for the other indicator.

Most (7 out of 9) themes for articles with low citation rates or ChatGPT scores applied to both categories (Table 4). These included research about university libraries, education, and open access publishing as well as survey-based studies and those specifying a geographic location (often associated with a survey). In contrast, whilst ChatGPT did not value management related research, it was not under-cited, and research about indexing had few citations but did not get particularly low scores from ChatGPT. Articles mentioning the term indexing were about citation indexes, indexing and abstraction services, and indexing as part of information retrieval algorithms.

Table 4. Themes found from WATA for title and abstract terms associated with low ChatGPT scores or low citation rates for LIS articles from 24 journals 2016-2020. Source: Author's own work.

Theme	**Statistically significant terms**	**ChatGPT**	**NLCS***
Management of libraries, resources or business	Management	***	
Surveys	Questionnaire, respondent, survey, descriptive	***	(***)
University libraries and resources	Library, libraries, university, librarian, staff, institution, service, training, professional, collection, patron, e-resource	***	***
Students and education	Student, education, faculty	***	***
Style: past tense and indirect language	Was, were, analyzed, included, revealed	***	***
Geographic locations	Nigeria, India, Indian, states	***	***
Open access publishing	Access	***	***
Books and reading	Book, print	(***)	***
Indexing	Title		***

*Three stars indicates at least one term statistically significant with $p<0.001$. No stars indicates that no term is statistically significant with $p<0.05$. Brackets indicate that the theme appears to be less prominent than for the other indicator.

A reviewer asked if words expressing positivity or certainty associated with high or low ChatGPT scores. Using the word association data underlying the WATA tables above, we searched for such words in the statistically significant $p<0.001$ terms, finding positive/certainty words that were more likely to be in the higher scoring subset: outperform (rank 33), demonstrate (36), state-of-the-art (76), contribution (122), substantial (141). We also found one negative or uncertainty words that were more likely to occur in the higher performing subset: argue (30), suggest (rank 42), negative (104), bias (126). In the subset with the lower ChatGPT scores, positive or certainty words in included showed (rank 33), recommended (62), highest (70), maximum (71) and there were no negative or uncertainty words with $p<0.001$ statistical significance. Overall, then there is not a clear association between positivity and/or certainty and higher or lower ChatGPT scores.

Discussion

The results of this article might have changed if different decisions had been made, such as another year range, a different cut-off for journal size, a different (or no) maximum number of articles per journal, or another citation database. For example, the Journal of Informetrics editorial board resigned in January 2019 (Larivière, 2019) and Quantitative Science Studies emerged to replace it, so a post-2019 analysis might give different overall results for these two in particular. Also, several stages in the research process involved subjective decisions, and especially the WATA analysis so the results have an element of subjectivity. Finally, the reason for the correlation is unknown. ChatGPT might have access to citation data, although LLMs are not ideal for analysing fine grained numerical information, or it might deduce the value of articles indirectly from how often they have been cited (although it is not given a

reference, only a title and abstract), or from online information about the reputation of the publishing journal. Thus, the results should be interpreted as a perspective on the value of ChatGPT for LIS research rather than definitive conclusions.

The journal-related results should not be interpreted as estimating the value of academic journals, but only the average quality of the articles that they publish. A journal's value to the academic community might include other factors, such as timeliness, relevance to the profession, and community support.

Comparison with prior work

For RQ1, the finding that ChatGPT scores tend to be higher for more cited LIS articles is novel, although it aligns with previous evidence from almost all broad fields, including the single REF grouping Communication, Cultural and Media Studies, Library and Information Management, that both citation rates and ChatGPT scores positively correlate with expert research quality scores (Thelwall & Yaghi, 2024). Previous research has also compared journal citation rates with average research quality scores, finding (Spearman) correlations between 0.03 and 0.5 for UK Research Excellence Framework (REF) articles in all fields, with the Spearman correlation for Communication, Cultural and Media Studies, Library and Information Management being 0.16 (Thelwall et al., 2023a). In this context, the journal level (Spearman) correlation between citation rates and mean ChatGPT scores reported above of 0.843 is very high, given that ChatGPT scores are indicators of research quality rather than direct measures of it. The high value could be a cumulative effect of the narrow LIS field, the wider quality range of the articles (REF submissions are self-selected for quality), and the exclusion of small journals from the LIS dataset (reducing statistical noise in the data). Alternatively, ChatGPT might be accessing and influenced by citation data or by noticing citations in texts.

For RQ2, previous studies have investigated highly cited LIS research from multiple perspectives. The most cited 5% of LIS research 2010-2014 included information seeking, information retrieval, bibliometrics, search engines, websites, and electronic government (Chang et al., 2015). The only overlap with the current paper is e-government, perhaps due to topic changes over time, the earlier paper not comparing high to lower cited studies, or the earlier paper including all of a journal's output rather than a sample of 250 articles. A follow-up study found mainly bibliometric topics (Hou et al., 2018), with no overlap with the current paper. A more recent study of publications from 1983 to 2018 found information systems research to dominate highly cited LIS papers (Kharabati-Neshin et al., 2021), again not overlapping with the current study. Finally, an investigation of the general Library & Information Science Research (LISR) journal 1994-2020 listed highly cited papers but did not classify their topics (Garg & Singh, 2022).

One previous study has used a version of WATA to investigate themes in articles with high or low REF quality scores from experts. The students and education theme aligns with a more general observation that in the UK REF, education-focused research tended to receive lower expert quality scores for arts and humanities research (although not specifically for the Communication, Cultural and Media Studies, Library and Information Management category), and that the first-person writing style associates with higher scores in many fields (Thelwall et al., 2023b).

Journal anomalies

There are no substantial journal outliers in terms of the relationship between ChatGPT and citation rates (Figure 2), although the *International Journal of Information Management*

seems to be relatively highly cited for its ChatGPT scores and the *Journal of Modern Information, Evidence Based Library and Information Practice*, and the *Proceedings of the Association for Information Science and Technology* all have relatively high ChatGPT scores for their citation rates.

The *Journal of Modern Information* is a Chinese language journal with translated titles and abstracts indexed in Scopus. It may attract relatively few Scopus-indexed citations for the quality of its work due to authors tending to publish in Scopus-indexed journals usually not speaking Chinese and not reading its articles, or Chinese-speaking citing authors tending to publish in Chinese language journals not indexed by Scopus.

Evidence Based Library and Information Practice is a platinum open access journal, which should give it a citation advantage, but seems not to, at least in comparison to ChatGPT. As suggested by its title and focus statement, "The purpose of the journal is to provide a forum for librarians and other information professionals to discover research that may contribute to decision making in professional practice" (Medaille, 2024) this journal has an applied orientation and it is possible that ChatGPT is better able to reflect its non-scholarly impacts (e.g., support for library practice) than are citation rates.

In the last case, the relatively few *Proceedings of the Association for Information Science and Technology* citations may be due to this volume being conference proceedings and its contributions (which include poster-papers) might often be developed into full papers that were subsequently published and would become the primary target for citations to the work. In summary, in all three anomalous cases, the ChatGPT score seems to be a more reasonable reflection of the average quality of the articles in a journal/serial than its citation rate, but this is a subjective conclusion.

Conclusion

The results show that ChatGPT quality scores for LIS research give similar results to citation rates at the journal level and in terms of most high (or low) ChatGPT/citation themes. Although ChatGPT may have some awareness of citations or citation data, this provides some evidence to validate both as research quality indicators, or at least is consistent with this hypothesis. The surprisingly high correlation at the journal level suggests that the citation indicator limitation of not directly reflecting originality, rigour and societal impact is not a major problem in LIS on a large scale. Nevertheless, the subjective analysis of anomalies suggests that ChatGPT may be marginally better than citation rates for a few unusual LIS journals. ChatGPT may tend to score article more highly if they explicitly mention novelty or research contexts, however, which may indicate a writing strategy bias.

The results also point to topics that do not tend to be highly cited or highly rated by ChatGPT, including academic libraries and education. Since these are both core tasks for the LIS field, the low scores may reflect academics conducting research on their practice for professional development purposes or for local insights, rather than as part of a developed programme of research. From this perspective, it may not be concerning that such research seems not to match the quality or scholarly impact of other studies in the field.

From an individual researcher perspective, if large language model scores start to be used alongside citations for research quality indicators then there is a risk that academics try to game the scores, perhaps by stuffing their abstracts with keywords associated with higher scores or by overclaiming rigour, originality and significance in their abstracts. Such a strategy might well be counterproductive, however, since misleading abstracts may well alienate the

core audience for researchers, which continues to be reviewers (in the first instance) and other academics.

From a research evaluation perspective, since ChatGPT is not evaluating research quality but only guessing, it does not threaten the hegemony of academic experts for the research evaluation task. Instead its role can be similar to that of bibliometrics, such as providing evidence to support human judgement when experts disagree, lack expertise or are uncertain. In this context, and in an era of apparently routine overclaiming for large language models, it is important that evaluators harnessing ChatGPT scores to support their judgements are cautioned that the scores are only guesses.

In terms of recommendations for library practice, academic librarians can use average ChatGPT scores for journals (Figure 2 for LIS; possibly similar figures in other academic publications) as a second opinion about the quality of journals for patrons concerned about where to submit their paper. In addition, they may wish to use ChatGPT's estimates based on article titles and abstracts for a quick opinion about the quality of a patron's paper as a form of formative feedback for them. To get the best results, the REF instructions for quality evaluation and the paper should be submitted at least 5 times to an LLM in separate chat sessions and the results averaged. The score should then be compared to other papers from the same field (e.g., Figure 1 if it is a LIS paper, otherwise search for similar published research about ChatGPT in other fields to norm reference against, or norm reference against other papers from the same scholar). Permission should first be sought from the author(s) for this for unpublished manuscript titles and abstracts since these are unlikely to be covered by copyright exemptions and the authors should be aware that their knowledge may be used by the LLM and recycled in response to other users' prompts.

In conclusion, ChatGPT seems to give insights into LIS research and a slightly different perspective to citation analysis for on which journals tend to publish the highest quality research. It therefore seems to be a useful additional tool for estimating the quality of publications in the field. Of course, the article-level scores are too unreliable to be replace human expert opinions and in any case would need ranking or scaling (e.g., Thelwall & Yaghi, 2024) to match the score ranges typically used by human experts.

References

Aksnes, D.W., Langfeldt, L., and Wouters, P. (2019), "Citations, citation indicators, and research quality: An overview of basic concepts and theories", *Sage Open*, Vol. 9 No. 1, p. 2158244019829575.

Benjamini, Y. and Hochberg, Y. (1995), "Controlling the false discovery rate: A practical and powerful approach to multiple testing", *Journal of the Royal Statistical Society: Series B*, Vol. 57 No. 1, pp. 289-300.

Chang, Y.W., Huang, M.H., and Lin, C.W. (2015), "Evolution of research subjects in library and information science based on keyword, bibliographical coupling, and co-citation analyses", *Scientometrics*, Vol. 105, pp. 2071-2087.

de Bellis, N. (2009), *Bibliometrics and citation analysis: from the science citation index to cybermetrics*. Scarecrow Press.

Figuerola, C. G., García Marco, F. J., & Pinto, M. (2017). "Mapping the evolution of library and information science (1978–2014) using topic modeling on LISA", Scientometrics, Vol. 112, pp. 1507-1535.

Garg, K.C. and Singh, R.K. (2022), "A bibliometric study of papers published in library and information science research during 1994-2020", *DESIDOC Journal of Library & Information Technology*, Vol. 42 No. 1, pp. 57-63. https://doi.org/10.14429/djlit.42.1.17480

Hicks, D., Wouters, P., Waltman, L., De Rijcke, S., and Rafols, I. (2015), "Bibliometrics: The Leiden Manifesto for research metrics", *Nature*, Vol. 520 No. 7548, pp. 429-431.

Hou, J., Yang, X., and Chen, C. (2018), "Emerging trends and new developments in information science: A document co-citation analysis (2009–2016)", *Scientometrics*, Vol. 115, pp. 869-892.

Kelly, J. (2024), "An introduction to copyright law and practice in education, and the concerns arising in the context of GenerativeAI", available at: https://nationalcentreforai.jiscinvolve.org/wp/2024/03/11/copyright-and-concerns-arising-around-generative-ai/ (accessed 23 November 2024).

Kharabati-Neshin, M., Yousefi, N., Mirezati, S.Z., and Saberi, M.K. (2021), "Highly cited papers in library and information science field in the Web of Science from 1983 to 2018: A bibliometric study", *Library Philosophy and Practice*, Vol. 2021, pp. 1-22.

Kousha, K., and Thelwall, M. (2024), "Artificial intelligence to support publishing and peer review: A summary and review" Learned Publishing, Vol. 37 No. 1, pp. 4-12.

Langfeldt, L., Nedeva, M., Sörlin, S., and Thomas, D.A. (2020), "Co-existing notions of research quality: A framework to study context-specific understandings of good research", *Minerva*, Vol. 58 No. 1, pp. 115-137.

Larivière, V. (2019), "Resignation of the editorial board of the Journal of Informetrics", available at: https://www.issi-society.org/blog/posts/2019/january/resignation-of-the-editorial-board-of-the-journal-of-informetrics/ (accessed 23 November 2024).

Liang, W., Izzo, Z., Zhang, Y., Lepp, H., Cao, H., Zhao, X., and Zou, J.Y. (2024a), "Monitoring AI-modified content at scale: A case study on the impact of ChatGPT on AI conference peer reviews", *arXiv preprint*, arXiv:2403.07183.

Liang, W., Zhang, Y., Cao, H., Wang, B., Ding, D.Y., Yang, X., and Zou, J. (2024b), "Can large language models provide useful feedback on research papers? A large-scale empirical analysis", *NEJM AI*, AIoa2400196.

MacRoberts, M.H. and MacRoberts, B.R. (2018), "The mismeasure of science: Citation analysis", *Journal of the Association for Information Science and Technology*, Vol. 69 No. 3, pp. 474-482.

Medaille, A. (2024), "About the journal", available at: https://journals.library.ualberta.ca/eblip/index.php/EBLIP/about (accessed 23 November 2024).

Merton, R.K. (1973), *The Sociology of Science: Theoretical and Empirical Investigations*, The University of Chicago Press, Chicago, IL.

Moed, H. F. (2006), "Citation analysis in research evaluation", Springer Science & Business Media.

OpenAI (2024a), "Data usage for consumer services FAQ", available at: https://help.openai.com/en/articles/7039943-data-usage-for-consumer-services-faq (accessed 3 November 2024).

OpenAI (2024b), "Consumer privacy at OpenAI", available at: https://openai.com/consumer-privacy/ (accessed 3 November 2024).

Rushforth, A. and Hammarfelt, B. (2023), "The rise of responsible metrics as a professional reform movement: A collective action frames account", *Quantitative Science Studies*, Vol. 4 No. 4, pp. 879-897.

Siddique, N., Rehman, S. U., Khan, M. A., & Altaf, A. (2021), "Library and information science research in Pakistan: A bibliometric analysis, 1957–2018", Journal of librarianship and information science, Vol. 53 No. 1, pp. 89-102.

Thelwall, M. (2017), "Three practical field normalised alternative indicator formulae for research evaluation", *Journal of Informetrics*, Vol. 11 No. 1, pp. 128-151.

Thelwall, M. (2023), "Word association thematic analysis: Insight discovery from the social web", *SN Computer Science*, Vol. 4 No. 6, p. 827.

Thelwall, M. (2024a), "Can ChatGPT evaluate research quality?", *Journal of Data and Information Science*, Vol. 9 No. 2, pp. 1-21. https://doi.org/10.2478/jdis-2024-0013

Thelwall, M., Jiang, X., and Bath, P.A. (2025), "Evaluating the quality of published medical research with ChatGPT", Information Processing & Management, Vol. 62 No. 4, paper no. 104123. https://doi.org/10.1016/j.ipm.2025.104123.

Thelwall, M., Kousha, K., Makita, M., Abdoli, M., Stuart, E., Wilson, P., and Levitt, J. (2023a), "In which fields do higher impact journals publish higher quality articles?", *Scientometrics*, Vol. 128, pp. 3915-3933. https://doi.org/10.1007/s11192-023-04735-0

Thelwall, M., Kousha, K., Abdoli, M., Stuart, E., Makita, M., Wilson, P., and Levitt, J. (2023b), "Terms in journal articles associating with high quality: Can qualitative research be world-leading?", *Journal of Documentation*, Vol. 79 No. 5, pp. 1110-1123.

Thelwall, M., Kousha, K., Stuart, E., Makita, M., Abdoli, M., Wilson, P. and Levitt, J. (2023c), "In which fields are citations indicators of research quality?", *Journal of the Association for Information Science and Technology*, Vol. 74 No. 8, pp. 941-953.

Thelwall, M., Kousha, K., Wilson, P. Makita, M., Abdoli, M., Stuart, E., Levitt, J., Knoth, P., and Cancellieri, M. (2023d). Predicting article quality scores with machine learning: The UK Research Excellence Framework. *Quantitative Science Studies*, Vol. 4 No. 2, pp. 547–573. https://doi.org/10.1162/qss_a_00258

Thelwall, M. and Kurt, Z. (2024), "Research evaluation with ChatGPT: Is it age, country, length, or field biased?", available at: https://arxiv.org/abs/2411.09768 (accessed 23 November 2024).

Thelwall, M. (2024b), "Evaluating research quality with large language models: An analysis of ChatGPT's effectiveness with different settings and inputs", *Journal of Data and Information Science*, Vol 10 No. 1, pp. 1-19. https://doi.org/10.2478/jdis-2025-0011

Thelwall, M. and Yaghi, A. (2024), "In which fields can ChatGPT detect journal article quality? An evaluation of REF2021 results", available at: https://arxiv.org/abs/2409.16695 (accessed 23 November 2024).

Urs, S.R. and Minhaj, M. (2023), "Evolution of data science and its education in iSchools: An impressionistic study using curriculum analysis", *Journal of the Association for Information Science and Technology*, Vol. 74 No. 6, pp. 606-622.

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*, Vol. 43, pp. 129-139.

Wang, J. (2013), "Citation time window choice for research impact evaluation", *Scientometrics*, Vol. 94 No. 3, pp. 851-872.

Wang, Z., Zhang, H., Chen, H., Feng, Y. and Ding, J. (2024), "Content-based quality evaluation of scientific papers using coarse feature and knowledge entity network." *Journal of King Saud University-Computer and Information Sciences,* Vol. 36, No. 6, 102119.

Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., Hill, S., and Johnson, B. (2015), *The metric tide: Independent review of the role of metrics in research assessment and management*.

Zhou, R., Chen, L., and Yu, K. (2024), "Is LLM a reliable reviewer? A comprehensive evaluation of LLM on automatic paper reviewing tasks", in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 9340-9351.

Appendix: ChatGPT system instructions (Thelwall & Yaghi, 2024)

These instructions are presented exactly as input to ChatGPT as its system instructions, except that bullet points, bold text and first line indentation have been used to improve readability for the readers of this article. This is an exact quote from a previous paper (Thelwall & Yaghi, 2024) but quotation marks have not been added to avoid confusion about what the instructions are.

> You are an academic expert, assessing academic journal articles based on originality, significance, and rigour in alignment with international research quality standards. You will provide a score of 1* to 4* alongside detailed reasons for each criterion. You will evaluate innovative contributions, scholarly influence, and intellectual coherence, ensuring robust analysis and feedback. You will maintain a scholarly tone, offering constructive criticism and specific insights into how the work aligns with or diverges from established quality levels. You will emphasize scientific rigour, contribution to knowledge, and applicability in various sectors, providing comprehensive evaluations and detailed explanations for its scoring.
>
> - **Originality** will be understood as the extent to which the output makes an important and innovative contribution to understanding and knowledge in the field. Research outputs that demonstrate originality may do one or more of the following: produce and interpret new empirical findings or new material; engage with new and/or complex problems; develop innovative research methods, methodologies and analytical techniques; show imaginative and creative scope; provide new arguments and/or new forms of expression, formal innovations, interpretations and/or insights; collect and engage with novel types of data;

and/or advance theory or the analysis of doctrine, policy or practice, and new forms of expression.

- **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.

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. Nor will there be an assumption that work published in a language other than English or Welsh is necessarily of a quality that is or is not internationally benchmarked.

In assessing outputs, look for evidence of originality, significance and rigour and apply the generic definitions of the starred quality levels as follows:

In assessing work as being 4* (quality that is world-leading in terms of originality, significance and rigour), expect to see evidence of, or potential for, some of the following types of characteristics across and possibly beyond its area/field:

- a primary or essential point of reference;
- of profound influence;
- instrumental in developing new thinking, practices, paradigms, policies or audiences;
- a major expansion of the range and the depth of research and its application;
- outstandingly novel, innovative and/or creative.

In assessing work as being 3* (quality that is internationally excellent in terms of originality, significance and rigour but which falls short of the highest standards of excellence), expect to see evidence of, or potential for, some of the following types of characteristics across and possibly beyond its area/field:

- an important point of reference;
- of considerable influence;
- a catalyst for, or important contribution to, new thinking, practices, paradigms, policies or audiences;
- a significant expansion of the range and the depth of research and its application;
- significantly novel or innovative or creative.

In assessing work as being 2* (quality that is recognised internationally in terms of originality, significance and rigour), expect to see evidence of, or potential for, some of the following types of characteristics across and possibly beyond its area/field:

- a recognised point of reference;
- of some influence;
- an incremental and cumulative advance on thinking, practices, paradigms, policies or audiences;
- a useful contribution to the range or depth of research and its application.

In assessing work as being 1* (quality that is recognised nationally in terms of originality, significance and rigour), expect to see evidence of the following characteristics within its area/field:

- an identifiable contribution to understanding without advancing existing paradigms of enquiry or practice;
- of minor influence.

(Thelwall & Yaghi, 2024)