**Article:**
Thelwall, M. (2026) ChatGPT estimates of the quality of published conference papers from their titles and abstracts. Data Technologies and Applications. ISSN: 2514-9288

https://doi.org/10.1108/DTA-03-2025-0205

# ChatGPT estimates of the quality of published conference papers from their titles and abstracts[1]

Mike Thelwall, University of Sheffield, UK.

**Purpose**: Although citation-based indicators are sometimes used to help evaluate the quality of papers in conference-based fields, conferences seem to be less systematically indexed in the major citation indexes, and the value of citation counts as research quality indicators for them is unknown. In response, this article investigates whether ChatGPT might provide a suitable alternative.
**Design/methodology/approach:** ChatGPT was used to assign a quality score to conference papers from 2020 from the 23 narrow fields that are partly conference-based in the sense of having at least 30% conference papers indexed in Scopus. The results were compared with citation counts and conference rankings.
**Findings**: ChatGPT research quality score predictions based on paper titles and abstracts alone correlated positively and statistically significantly with citation rates in all fields at the paper level and mostly at the conference level. Expert-based citation-informed conference rankings conformed slightly more closely with geometric mean citation rates than with mean ChatGPT scores.
**Research limitations/implications:** No direct measure of paper research quality was used.
**Originality/value:** Whilst the evidence tends to support the value of both citations and ChatGPT as research quality indicators, it suggests that citations may be better, at least for older research, and that ChatGPT is a reasonable alternative for research that is too new to have attracted many citations, at least in conference-based fields.
**Keywords**: Research evaluation, scientometrics, bibliometrics, large language models, ChatGPT.

## Introduction

Research quality evaluation is a labour-intensive task but necessary to support appointment and promotion decisions, as well as many other aspects of academic life. It may be particularly problematic for those lacking relevant specialty expertise or time for an evaluation. In such contexts, quantitative indicators like citations or rejection rates might be harnessed to help decide the quality of a paper or conference series. Both are limited, however. Whilst citations do not directly reflect the societal impacts, rigour and originality of a study (Aksnes et al., 2019) and take years to accumulate in sufficient numbers (Wang, 2013), rejection rates might be influenced by the size of a conference venue as well as gaming by the programme committee. Thus, an alternative is needed.

Whilst machine learning and modelling has been used for quality or citation prediction within limited success for journal articles (Kousha & Thelwall, 2024), a new possible solution is to ask a Large Language Model (LLM) like ChatGPT to provide an evaluation of a paper or conference of interest, given that is works well for many natural language processing tasks (Kocoń et al., 2023). If it worked for conference paper evaluation, this would also overcome

[1] Thelwall, M. (2026). ChatGPT estimates of the quality of published conference papers from their titles and abstracts. *Data Technologies and Applications*.

the inability to use citations for just-published papers. This approach has already been shown to work to some extent for one conference (Thelwall & Yaghi, 2025b) and many journals and fields (Thelwall & Yaghi, 2025a) and has been investigated for conference peer review (Liang et al., 2024b). Of course, LLMs are also used by reviewers to help write their reports (Liang et al., 2024a). In this context, a systematic evaluation is needed to check whether the use of ChatGPT is more generally applicable to conferences.

This article assesses whether ChatGPT research quality scores for conference papers and conferences are a plausible alternative to citation rates, as a research quality indicator. The goal is not to judge whether ChatGPT can *evaluate* research but only whether it can provide *useful* research quality score estimates. As part of this, the system will only process titles and abstracts to get the estimates since these are readily available for conferences, whereas full texts sometimes are not. Unfortunately, there is no gold standard set of conference papers with research quality scores except in isolated open peer review special cases (e.g., Kang et al., 2018), so an indirect approach was taken for the primary tests: correlating ChatGPT scores with citation rates. A positive result would tend to support the value of both as research quality indicators, without proving that either is valid. Whilst a positive result would leave open the possibility that both are invalid, as a minimum it would reduce the likelihood of this being true, in the absence of other information. The third research question applies to a higher aggregation level, entire fields, to identify whether there is an overall pattern in the results The following research questions drive the study. The final one is the most important but there is relatively little evidence to answer it. Note that the questions all apply only to conference-based fields, which are a minority in academia.

- RQ1: Do more cited conference papers tend to get higher ChatGPT scores in any or all conference-based fields?
- RQ2: Do conferences with more cited papers tend to publish papers with higher ChatGPT scores in any or all conference-based fields?
- RQ3: Do conference-based fields with more cited papers tend to publish papers with higher ChatGPT scores?
- RQ4: Do citation rates or ChatGPT scores better reflect expert conference rankings in conference-based fields?

# Background

There has been surprisingly little research into conference-based fields as a phenomenon, and the reasons why specialities have evolved around conferences rather than journals has not been investigated systematically. In addition, with one minor exception (Aduku et al., 2017), there seems to have been no research into any aspect of research evaluation for engineering conferences outside of computer science. This section therefore contextualises conferences within scholarly communication to give a broad perspective for this research.

## *Conferences and scholarly communication*

Although originally communicated orally and visually (showing and telling), with the invention of writing in Mesopotamia (Iraq) 5000 years ago, human knowledge is now frequently transmitted primarily through text. In the context of organised knowledge creation, acquisition and transmission, the standard mechanisms have included clay tablets, papyrus, parchment, and books. The modern era for science is usually dated to the seventeenth century, when Western “philosophers” had absorbed much of the learning of the Greek,

Roman and Muslim worlds and started to generate substantial bodies of new understandings, including with the new scientific method of discovery (Henry, 2008; Morus, 2022). In this era, knowledge was originally exchanged and debated through letters between scholars and public or private meetings of interested parties, increasingly organised through membership-based bodies or researchers like the Royal Society in the UK. These invented the modern academic journal, which originally published letters of new findings or arguments, as well as transcripts of talks at meetings (Oppenheim et al., 2000). The earlier invention of the printing press allowed these journals to effectively replace letters between scholars as well as disseminating knowledge to people that were unable to attend meetings (Greco, 2024).

From the nineteenth century, the number of academic journals grew and specialised exponentially (e.g., Mabe & Amin, 2001). In parallel, public meetings of interested scholars evolved into academic conferences focusing on specialisms and sharing their proceedings in dedicated volumes, although sometimes still in journal special issues. In the second half of the 20th century, academic conferences became widespread, facilitated by transportation and communication technologies, to support networking, exchanges of ideas and build the cohesion of fields (Agar, 2012; de Leon & McQuillin, 2020; Hauss, 2021). In a few cases, large conferences were also created to coordinate large scale big science projects and to promote international collaboration (Agar, 2012). Now, however, environmental issues and technologies to support online participation are causing a rethink of in-person events (Raby & Madden, 2021).

## *Advantages of conferences*

A key advantage of conferences seems to be speed of dissemination. Although this has been undermined by the increasing acceptance of pre-printing, a scholar submitting to a conference can expect to have their work ingested by interested parties in their field within six months of the initial submission. Whilst a journal submission could easily be delayed by a year for reviewing and another for publishing (Björk & Solomon, 2013) (although online first/early view publishing now provide alternatives), a conference has a fixed deadline, so a publish/present decision is guaranteed by the conference start date. Thus, for this reason, conferences seem particularly suited to fast moving fields. This situation seems to have changed to some extent now. Whilst there have been some quick alternatives for a long time, such as short form “letters” journals in some fields, and prestigious journals that are able to almost guarantee fast peer review, instant publishing (although not instant audience) is now guaranteed by preprints and quick publishing in journals is widely available through publishers that prioritise this, such as MDPI, or through publish-before-review platforms, like F1000Research.

There is also evidence that the dynamics of fields can be shaped by conference interactions (Henderson, 2015; Recker et al., 2016). This may be an advantage to those that gain from the reshaping and perhaps the whole field.

## *Field differences in the role of conferences*

The role of conferences varies substantially between fields. In some cases, scholars submit short abstracts about their ideas and then give talks to small groups of people who are interested in their topic. These seem to serve primarily for testing preliminary ideas or sharing results. For example, 37% of abstracts presented at biomedical conferences are subsequently published as journal articles (Scherer et al., 2018). At the other extreme of importance, conference proceedings form the main knowledge record and exchange medium for some

fields, such as computational linguistics. For these, most scholars may view academic journals as secondary.

Other field-specific advantages of conferences include the ability to demonstrate physical artefacts (e.g. robots, paintings, architecture models) and to interact informally with other specialists and industrial experts (e.g., Wang et al., 2017), which may be important for collaborative fields and when industrial sponsorship is important. Historically, displaying videos and colour images may also have been only possible at conferences. In addition, some fields may be conference-based for historical reasons, perhaps because some of the above issues were important in the past even if they are no longer important now.

The importance of conferences in computer science has been previously demonstrated with evidence that the citation rates of some computer science conferences are higher than those for most journals (Freyne et al., 2010; Kochetkov et al., 2010; Vrettas & Sanderson, 2015). Even within computer science, that some specialities have a greater conference focus than others, however (Kim, 2019).

Although all conferences are unique, the following illustrate the origins of some important examples, highlighting the field specific issues that led to their formation.

- The TREC (Text REtrieval Conference) was started in 1992, initiated by the U.S. National Institute of Standards and Technology (NIST) and the Defense Advanced Research Projects Agency (DARPA) to promote natural language processing (NLP). These departments provided the funding/resources to design a series of language processing tasks and generate data for these tasks. Researchers were then challenged to solve these tasks and, if they registered for the conference, were sent a sample of the data. By the time of the conference applicants had to design a program to solve the task and submit a conference paper describing that task. At the conference, all systems would be tested no new data and the winner declared. For the organisers, this was an effective way to generate free labour on a complex task by stimulating academic competition. For the participants, this was an opportunity to test their theories, ideas, and software development skills on ready made datasets, which would otherwise require substantial resources to develop and validate (Voorhees & Harman, 1999). The nature of the competitions meant that an in-person event was the only practical solution.
- The Association for Computational Linguistics (ACL) conference series began in 1963 (with a different name in the USA. It is now arguably the most prestigious NLP conference series and more important than all NLP journals. Its prestige may stem from its early origins, whereas the importance of NLP conferences may be due to the fast evolution of computing technology (hardware and software), making journals historically too slow.
- The IEEE Engineering in Medicine and Biology Society (EMBS) conference series has been held since 1960 and, despite its name was founded by electronic engineers (https://www.embs.org/about/history/), so the attraction of this conference relative to journals might be fast-moving computing technology development.
- The European Safety and Reliability Conference (ESREL) conference series began in 1989, or 1992 under its current name. One of its stated aims is to bring together researchers and practitioners (esrel2025.com), which may be particularly important in this field; this networking is not possible through journals.
- The International Astronautical Congress (IAC) began in 1950, and its role includes bringing together multidisciplinary space science researchers with industry

professionals and the public (https://www.iafastro.org/events/iac/). Its conference-specific advantages may be multidisciplinary networking in a technology sector for which technological coordination is important, and perhaps also fast-moving technological change.

- The International Conference on Industrial Engineering and Operations Management (IEOM) conference series began in 2010 (https://ieomsociety.org/conferences/) and one of its aims is to foster networking between researchers and practitioners (https://ieomsociety.org/singapore2025/).

These cases illustrate the relevance of conferences to academics in the context of fast technological change, networking, and access to resources for specialists in areas where these are particularly important.

# Methods

As suggested by the research questions, the research design was to create a large dataset of conference papers with citation counts in multiple conference-based fields, to score them with ChatGPT, and then to correlate the results with citation rates. The same data is correlated with expert rankings of conferences, when available, for RQ3 (see below).

## *Field selection*

The concept of conference-based fields is vague but reflects the observation that conferences are central to research publishing in some disciplines, but books, journal articles or artworks dominate others. For simplicity, fields were equated with Scopus All Science Journal Classification (ASJC) codes, which seem narrow enough to capture a field, which is itself an imprecise concept (Sugimoto & Weingart, 2015).

A field was considered moderately or mainly conference-based if at least 30% of its Scopus-indexed documents in 2020 were conference papers. This threshold was selected subjectively after examining a list of fields in Scopus ordered by proportion of conference papers. This produced 23 fields (for a list, see the tables and figures in the results). Of course, a field may be journal-based with many irrelevant conferences (perhaps abstract-only) but these are unlikely to be indexed by Scopus since it considers citation rates in its decision to index publications and only indexes full text conferences (Elsevier, 2024).

## *Conference paper metadata downloading and cleaning*

For each field, the conference papers in 2020 were downloaded from Scopus in November 2024 using its Applications Programming Interface (API) with queries of the form:
PUBYEAR IS 2020 AND DOCTYPE(cp) AND SUBJMAIN(1802)
where cp means conference paper and 1802 is a field code. The fields in this dataset were Cited By (citation counts), Title, and Description (abstracts). The year 2020 was selected to give mature citation counts, with three years usually being sufficient for journals (Wang, 2013) and conferences seeming to move faster.

A small amount of data cleaning was necessary to remove duplicate papers. This included the same conference recorded under different names and the same paper recorded multiple times for a single conference, sometimes with minor typographical variations.

Some conference papers had short or no abstracts, apparently at the author's choice or because the contribution was not a full paper. Since ChatGPT needs an abstract to make reasonable estimate and non-paper contributions (e.g., posters, workshop summaries) would

add noise to the data, a 567-character minimum threshold was set to eliminate papers with the shortest 10% of abstracts as a simple heuristic to deal with this issue. Descriptive statistics for the final dataset are available in the online appendix (https://doi.org/10.6084/m9.figshare.28090556).

## *Conference selection*

Since one of the purposes of this article is to examine the conference level relationship between ChatGPT scores, conferences with few papers would have unreliable average scores and citation rates, and including large numbers of papers for big conferences would be statistically redundant. Hence, only conferences with at least 200 papers were included and a maximum of 1000 papers were selected for each conference, using a random number generator for selection when there were more. Both thresholds were chosen heuristically before the ChatGPT scores were collected. A minimum sample size was necessary for a conference because statistics calculated for conferences with few papers would be unreliable and obscure larger patterns. A maximum threshold was needed because the ChatGPT queries have to be paid for and, for example, increasing the threshold to 2000 would double the cost with little increase in accuracy. Since this is the first research of its type, it was not possible to conduct a power calculation to identify the two thresholds more systematically.

This process left 222 different conferences (some in multiple fields) from 23 narrow fields within five broad fields. The dataset was dominated by Computer Science and Engineering (Table 1).

Table 1. Sample statistics for fields with at least 30% conference papers in Scopus in 2020. Scopus narrow fields are listed under their associated Scopus broad field. Source: Author's own work.

| Field | ASJC | Papers | Conferences |
|---|---|---|---|
| **Computer Science** | | | |
| Artificial Intelligence | 1702 | 20771 | 59 |
| Computational Theory and Mathematics | 1703 | 3553 | 7 |
| Computer Graphics and Computer-Aided Design | 1704 | 2998 | 6 |
| Computer Networks and Communications | 1705 | 22035 | 67 |
| Computer Science Applications | 1706 | 23075 | 62 |
| Computer Vision and Pattern Recognition | 1707 | 10882 | 27 |
| Hardware and Architecture | 1708 | 7914 | 24 |
| Human-Computer Interaction | 1709 | 6511 | 13 |
| Information Systems | 1710 | 7565 | 20 |
| Signal Processing | 1711 | 13871 | 35 |
| Software | 1712 | 14263 | 28 |
| **Decision Sciences** | | | |
| Information Systems and Management | 1802 | 9502 | 29 |
| **Energy** | | | |
| Energy Engineering and Power Technology | 2102 | 15284 | 40 |
| **Engineering** | | | |
| Aerospace Engineering | 2202 | 5009 | 11 |
| Automotive Engineering | 2203 | 3003 | 7 |
| Control and Systems Engineering | 2207 | 12964 | 26 |
| Electrical and Electronic Engineering | 2208 | 29209 | 77 |
| Industrial and Manufacturing Engineering | 2209 | 6360 | 19 |
| Safety, Risk, Reliability and Quality | 2213 | 10580 | 33 |
| Media Technology | 2214 | 1121 | 4 |
| **Mathematics** | | | |
| Control and Optimization | 2606 | 15439 | 41 |
| Modeling and Simulation | 2611 | 6242 | 14 |
| Theoretical Computer Science | 2614 | 1739 | 5 |
| **Overall** | | **91189** | **222** |

## *ChatGPT 4o-mini scores*

Each conference paper in the sample was submitted to ChatGPT 4o-mini five times via its API (batch mode), accompanied by system instructions and the prompt, "Score this paper:", between 29 November and 5 December 2024. The code used is available online (https://github.com/MikeThelwall/LargeLanguageModels). The system instructions (see Appendix 1: https://doi.org/10.6084/m9.figshare.28327970.v1) were a definition of the task and scoring scale, taken from the engineering and physical sciences quality scoring guidelines from the UK Research Excellence Framework (REF) 2021 guidelines for assessors (Main Panel B: REF2021, 2019), as previously used for journal articles (Thelwall & Yaghi 2025a). The term "journal article" in the system instructions was not changed to the more accurate "conference paper" to avoid potentially confusing ChatGPT into thinking that it was not scoring a full academic output (since many conferences are abstract only and it was only fed with titles and abstracts) The guidelines are an appropriate source because they are intended to guide

quality scoring of academic research in the fields concerned, cover conference papers, and are somewhat discipline sensitive (there are three other guidelines for different fields). Although most outputs submitted to REF2021 were journal articles, a substantial minority of conference papers were submitted in some fields and the instructions used here were designed to be inclusive of them, so are appropriate. Full conference papers and journal articles are similar document types, with the main difference being the time-sensitive nature of the reviewing process although conference topics probably also tend to be more time-sensitive.

This type of source is appropriate because ChatGPT from version 3.5 onwards was developed to work with free text natural language instructions (Ouyang et al., 2022), as used by the originals of the guidelines. Whilst there are other concepts of research quality, rigour, originality and significance are usually accepted to but the three core components (Langfeldt et al., 2020) and these are the main criteria in the guidelines. Finally, they are in English, as are nearly all the conference paper abstracts and titles, which avoids an unnecessary mismatch for the ChatGPT processing.

The REF guidelines used for the system instructions do not differentiate between journal articles, conference papers or other output types. The prompting strategy is therefore essentially a request to ChatGPT to provide a quality score from 1* (nationally relevant) to 4* (world leading) for a conference paper based on its title and abstract. Although in some cases ChatGPT refused the request, citing insufficient information to decide, this was rare, and all papers received a score at least once. See Appendices 2 to 4 ( https://doi.org/10.6084/m9.figshare.28327970.v1) for example reports, slightly redacted to avoid drawing attention to the papers assessed.

ChatGPT's scores were extracted from its reports using a program written for this task (Webometric Analyst, AI menu, ChatGPT/Gemini extract REF scores, available at: github.com/MikeThelwall/Webometric_Analyst). Since the reports are not simple scores but several paragraphs of text justifying the scores, a set of heuristics (see online appendix: https://doi.org/10.6084/m9.figshare.28090556) was necessary to extract the scores from the texts. These were built into Webometric Analyst. For example, one of the heuristics was, to extract a score between 1* and 4* when it was the only non-space text between the strings, "Overall Average Score**:" and "***". When the Webometric Analyst score extraction rules failed, the ChatGPT report was shown to the author for a human evaluation. When ChatGPT did not give a score, a missing value was recorded. If it reported separate scores for originality, rigour and significance but not an overall score then the mean of these three scores was used instead, rounded to three decimal places. ChatGPT sometimes included this average in its reports, calculated to two decimal places. In three cases, the ChatGPT output had to be reinterpreted as a score (three black stars was interpreted as 3* [twice]; "10/12 overall "was interpreted as 3.333* given that the report included scores of 3*, 3* and 4* for originality, rigour, and significance). The final score for each paper was the average of five iterations because averaging has been shown to give better results (Thelwall, 2025). The papers were submitted in a random order (with a random number generator) and the entire set was repeated a further four times.

## *Citation rates*

For each paper, the citation count reported by Scopus was used. The three major citation databases, Scopus, the Web of Science and Dimensions.ai have broadly similar coverage of academic papers, so the choice of Scopus, the largest of the three (Martin Martin et al., 2021),

is unlikely to make much difference. Whilst Google Scholar probably indexes more conference papers than these, it lacks an API that would make data collection practical. Microsoft Academic might also have had larger coverage but is now defunct. For each conference, the geometric mean citation count was calculated in preference to the arithmetic mean because citation data is highly skewed. The geometric mean is closer to the median, in theory, and is much less influenced by individual highly cited papers, so is a better measure of central tendency.

Although field and year normalised citation counts are often used instead of raw citation counts, this was not necessary since all papers were indexed in the same year and only papers from the same narrow field were correlated against each other. The only advantage of field normalisation would be to consider the fact that some conferences are categorised in multiple fields, so these would have their citation rates increased in some of these fields and decreased in others. this would complicate the analysis and there is not a straightforward way to take into account the extent to which each conference matches the fields that it has been assigned to.

## *Correlations*

Mean ChatGPT scores were correlated against citation rates (calculated as above) to assess the extent to which they reflect similar phenomena. Since neither is a recognised quality indicator, positive correlations would give mutually supportive weak evidence for both being this. The strength of the correlations would also reveal the extent to which the results (rank orders) would be altered by a switch from one to the other. Spearman correlations were used since not all of the data is normally distributed. In particular, citation data is typically skewed, so Pearson correlations would be inappropriate. Skewing is irrelevant to the Sperman correlation because it uses only the rank orders from the correlated datasets: it is non-parametric. Bootstrapping (in R) was used for confidence intervals since there is no formula for this for Spearman correlations. Again, skewing is irrelevant because the bootstrapping only uses the ranks of the two datasets and not the raw data. Bootstrapping gives reasonable Spearman confidence intervals even for much smaller datasets than the one analysed here (Haukoos & Lewis, 2005; Ruscio, 2008).

## *Comparison with conference rankings*

Unfortunately, there are no public datasets of quality scores for conference papers that would allow direct paper-level analyses of the accuracy of ChatGPT, and it is not practical to assess them all for this paper because it is too time consuming. For example, 1120 field experts in REF2021 scored 185594 outputs over a year, whereas there are 91189 papers analysed in the current article so at the same rate it would require 550 experts for a year (at REF2029 rates, this would cost £5.5m, although the assessors also have other tasks). A weaker assessment would risk non-experts or brief evaluations that might mimic ChatGPT's lack of depth. Previous studies of ChatGPT have either used small samples of individually scored articles (<100) (e.g., Thelwall, 2025), or larger samples with research quality proxies derived from the UK REF2021 (e.g., Thelwall & Yaghi, 2025a). The latter is not possible here because the conference papers analyses are largely not from REF2021. Instead, public lists of quality scores for entire conferences were used to allow conference-level analyses.

Four public sets of conference rankings were found with online searches and used to compare with average ChatGPT scores and citation rates. The Australian Computing Research & Education (CORE, now International CORE) website ranks computer science conferences

(portal.core.edu.au/conf-ranks), and the Polish (www.gov.pl/web/nauka/komunikat-ministra-nauki-z-dnia-05-stycznia-2024-r-w-sprawie-wykazu-czasopism-naukowych-i-recenzowanych-materialow-z-konferencji-miedzynarodowych), Finnish (julkaisufoorumi.fi/en/publication-forum), and Norwegian (kanalregister.hkdir.no/sok) national publication multidisciplinary rankings include conferences although they mainly list journals.

All four lists are expert rankings but informed by citation data so the extent to which citation rates influence them is unknown. Nevertheless, these provide an opportunity to assess whether ChatGPT rankings or citation rate rankings fit existing expert rankings better. All the rankings are partial and so the correlations calculated as above relate to only conferences with rankings. To avoid selection biases, correlations are only compared on the same sets of conferences.

# Results

The results for the first two research questions are reported together since they naturally match.

## *RQ1, 2: Comparisons between citation rates and mean ChatGPT scores for papers and conferences*

Spearman correlations between mean ChatGPT scores and citation counts for papers are statistically significantly positive (95% confidence intervals exclude 0) in all 23 conference-based fields tested. In contrast, whilst conference-level correlations tend to be positive, there are three exceptions and more where the difference is not statistically significant (Figure 1). For conferences, the negative correlations are plausibly an artefact of small sample sizes (and hence wide confidence limits) rather than an underlying conference-based negative relationship. In both cases (papers and conferences), the overall trend is for positive correlations.

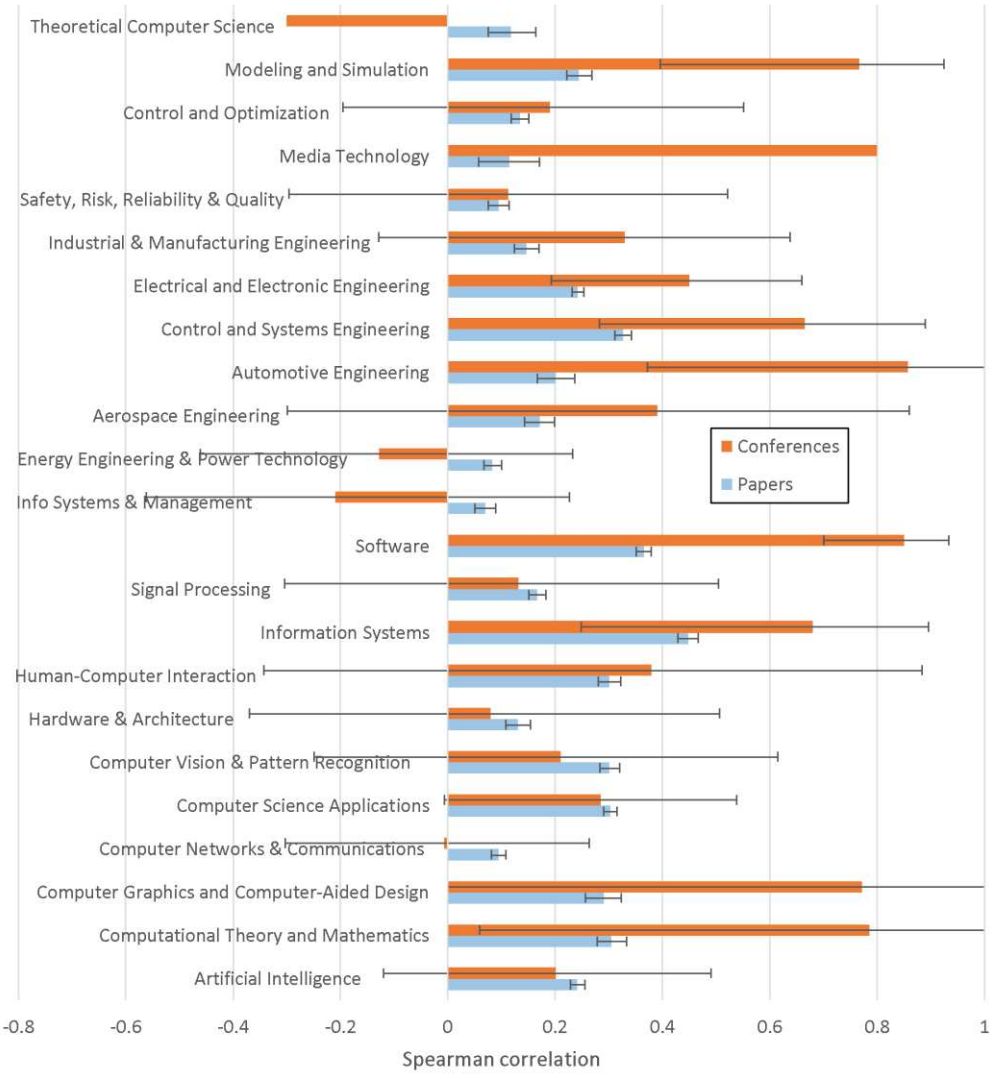


Figure 1. Spearman correlations between paper or conference (mean) GPT scores and citation counts (papers) or geometric mean citation counts (conferences) and bootstrapped 95% confidence intervals. Where error bars are missing, no confidence intervals could be calculated. Source: Author's own work.

As scatter graphs of the fields with the most conferences suggest, there is not a strong relationship between the conferences with the highest average ChatGPT score and the conferences with the highest citation rates within individual fields. A partial exception is that some computing-related fields have one or more conferences with the highest citation rate and highest mean ChatGPT score. For example, the anomaly in Computer Networks and Communications is the prestigious World Wide Web conference (Figure 2).


Figure 2. Scatter graphs of geometric mean citations per paper against arithmetic mean ChatGPT scores for conferences in the six fields with the most conferences. Source: Author's own work.

Using Artificial Intelligence as an example, despite averaging 5 ChatGPT scores per paper, by far the most common scores are 3* and 2*, especially for papers with few citations (Figure 3). The reason for the small positive Spearman correlation between ChatGPT scores and citation counts for conference papers within a field seems to be that papers with high ChatGPT scores (e.g., above 3.5*) are more likely to have many citations (e.g., above 25, or above 200). In

particular, despite there being few articles scoring 3.5* or higher (compared to the huge ball at 3* for example, in Figure 3) there are relatively many dots above the 200 line.

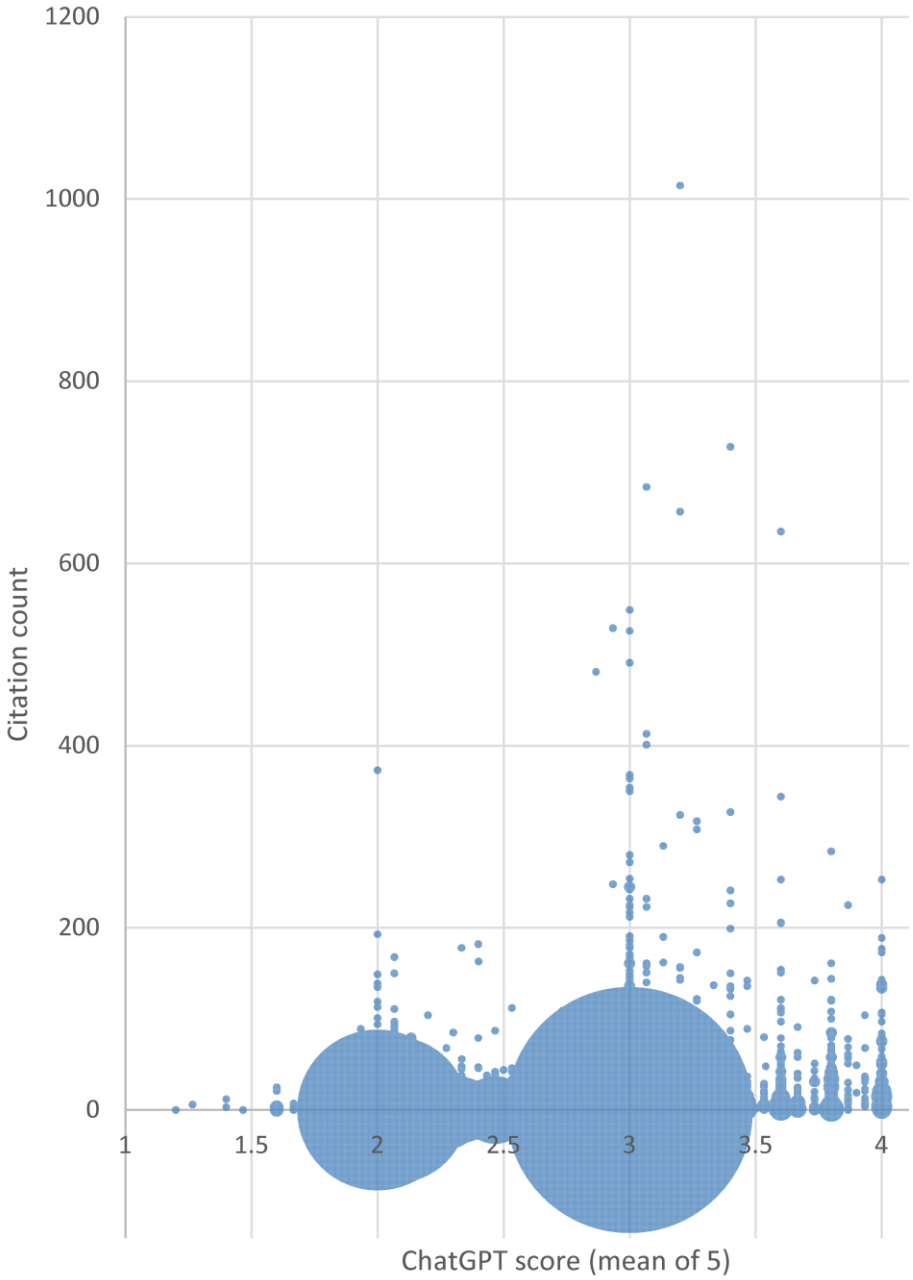


Figure 3. Bubble plot of citation ChatGPT score (mean of 5) for conference papers in 2020 from the Artificial Intelligence Scopus narrow field. Source: Author's own work.

## *RQ3: Cross-field comparisons between citation rates and mean ChatGPT scores*

In contrast to the situation for conferences within fields, there is a strong tendency for Scopus narrow fields with higher mean ChatGPT conference paper ChatGPT scores to have higher average (geometric mean) citation counts (Pearson's r=0.814, n=23; Pearson correlations were used because the trend is close to linear) (Figure 4). This may be due to more theoretical fields tending to attract more citations and higher ChatGPT scores, although it is not clear why ChatGPT could be more influenced by theory than by rigour and significance since its system

instructions (Appendix 1: https://doi.org/10.6084/m9.figshare.28327970.v1) tell it to consider all three.

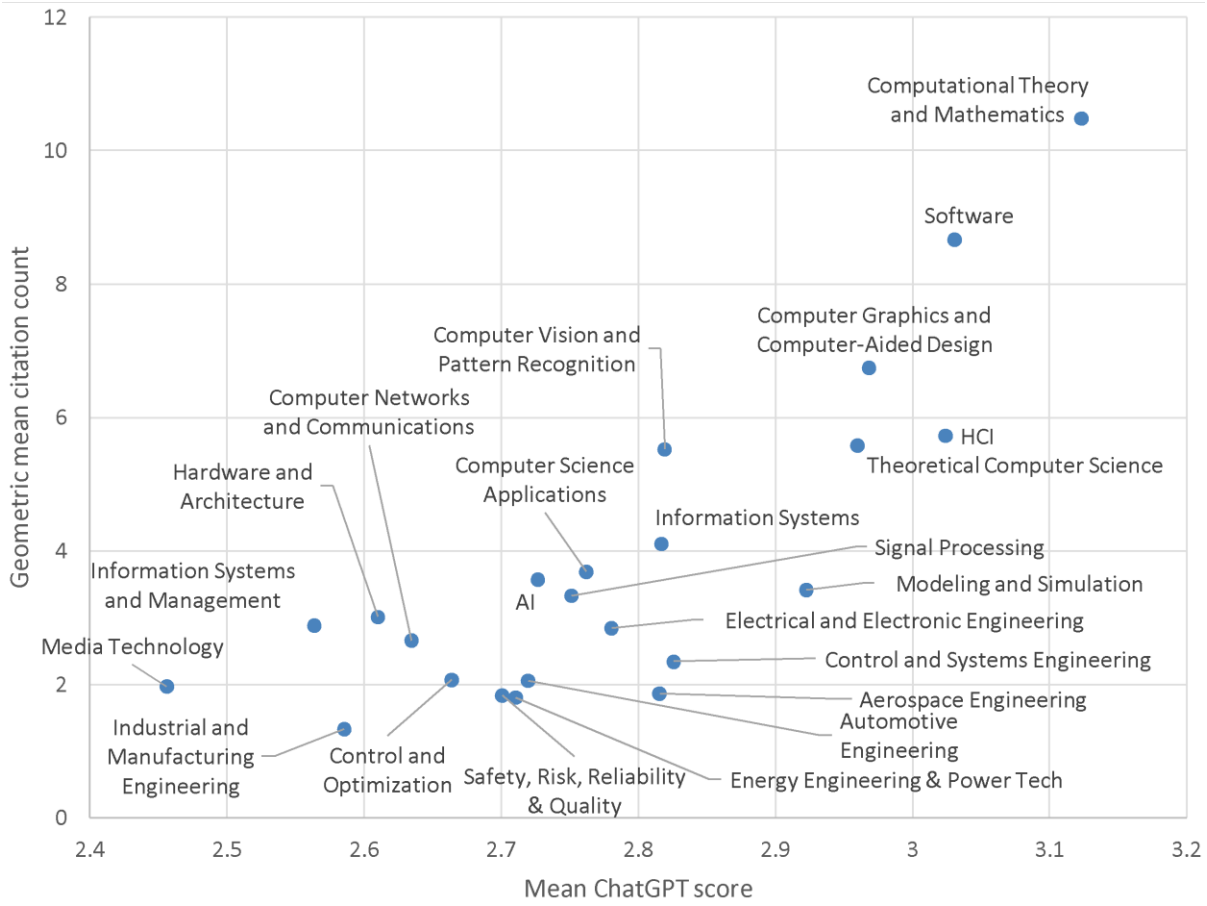


Figure 4. Geometric mean citations per paper against arithmetic mean ChatGPT score for Scopus narrow fields with at least 30% conference papers in 2020. Source: Author's own work.

## *RQ4: ChatGPT and citation comparisons with conference rankings*

The ranking comparisons were restricted to fields with at least 10 ranked conferences since small sample sizes would give unreliable correlations. Although the remaining numbers of ranked conferences per field were small, both mean ChatGPT score per paper and geometric mean citation count per paper correlated positively with expert ranks in most fields (e.g., Figures 5, 6, 7; for Norway only one field had at least 10 conferences, Electrical and Electronic Engineering: ChatGPT correlation: 0.46, Citation correlation: 0.35). The main exception is the Energy Engineering and Power Technology field (Figure 6).

None of the differences between the ChatGPT and citation correlations were statistically significant, but for three out of the four rankings (the three with the most data), the correlations overall tended to be higher for citations than for ChatGPT (Table 2) although the differences are typically not large (e.g., Figures 5, 6, 7; for Norway only one field had at least 10 conferences, Electrical and Electronic Engineering: ChatGPT correlation: 0.46, Citation correlation: 0.35). Thus, there is weak evidence to suggest that citation-based rankings align more closely with expert rankings than do ChatGPT-based rankings of conferences in fields for which conferences are important.

Table 2. Sample size weighted average Spearman correlations between expert conference ranks and mean ChatGPT scores or geometric mean conference citations. Source: Author's own work.

| Ranking | ChatGPT vs. Ranks | Citations vs. Ranks | Fields | Conferences |
|---|---|---|---|---|
| CORE | 0.54 | 0.78 | 16 | 138 |
| Finland | 0.48 | 0.59 | 18 | 216 |
| Norway | 0.41 | 0.32 | 7 | 51 |
| Poland | 0.47 | 0.53 | 17 | 175 |

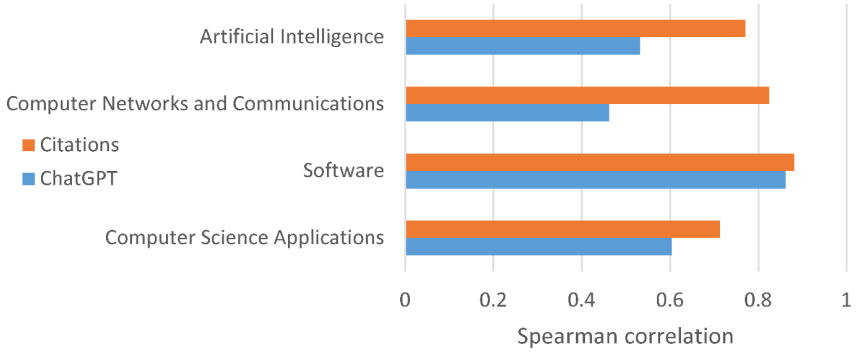


Figure 5. Spearman correlation between CORE conference rankings and geometric mean citation counts or mean ChatGPT scores by field (qualification: >9 conferences). Source: Author's own work.

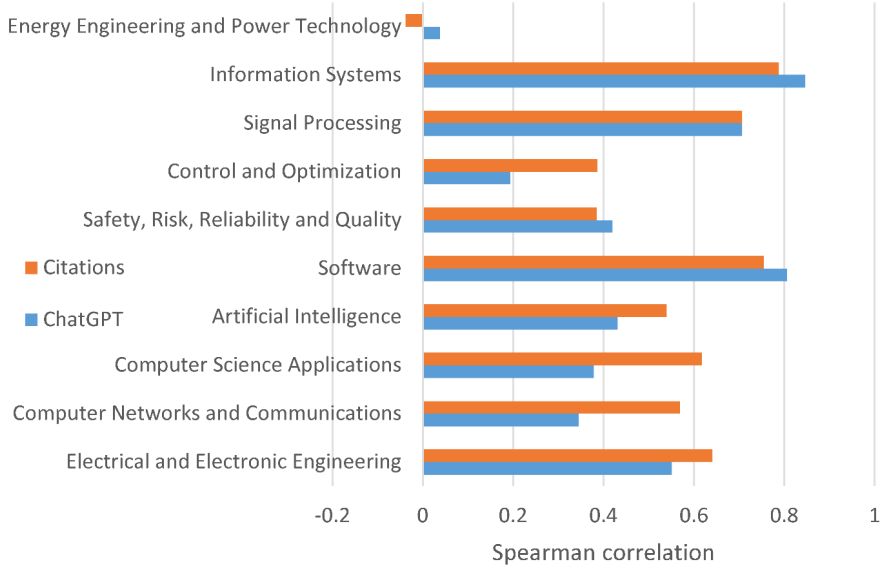


Figure 6. Spearman correlation between Finnish conference rankings and geometric mean citation counts or mean ChatGPT scores by field (qualification: >9 conferences). Source: Author's own work.

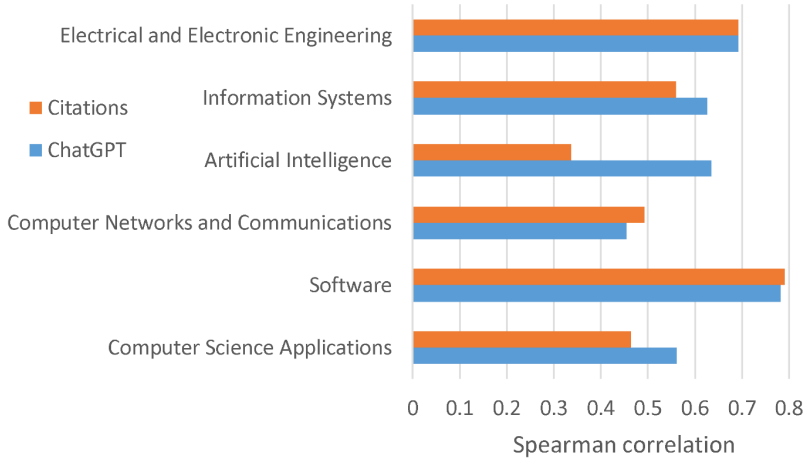


Figure 7. Spearman correlation between Polish conference rankings and geometric mean citation counts or mean ChatGPT scores by field (qualification: >9 conferences). Source: Author's own work.

## Discussion

This article is limited by the relatively arbitrary 30% conference papers threshold to select narrow fields to analyse, and the restriction to medium or large conferences in these fields. Different results may have been obtained from other citation databases and those that index abstract-only events. ChatGPT may also have "cheated" by taking into account online information, such as conference rankings or citation data, although this tends to be available in spreadsheets rather than textual formats that it would probably find easier to ingest/interpret. Although the prompt used did not warn ChatGPT against exploiting such sources, it is not clear that such a warning would have been heeded. Another limitation is that there might be types of conference contribution other than full primary research papers, and this would undermine the correlations. It is possible that abstract structures, field specific language, or paper types (e.g., theoretical/empirical) influence the correlations, but individual paper level quality scores might be needed to investigate this.

At the paper level, it seems likely that there is universally a positive correlation overall between paper quality and citation rates for fields in which conferences are important, since all correlations are positive for this aggregation level. Whilst this is mutual evidence that both paper citation rates and paper ChatGPT scores are at least weak indicators of research quality it does not prove this and it does not suggest which is better. Moreover, it is possible that this relationship does not exist for all conferences but only in terms of the difference between a few prestigious conferences and others. No previous research seems to have investigated the relationship between citation counts and either research quality or ChatGPT scores, so it is not possible to contextualise the results against prior related evidence. It is intuitively plausible that higher ChatGPT scores are more likely to align with higher quality research since they take into account non-academic impacts, but the field level analysis (albeit without specific evidence) suggests that it might tend to favour more theoretical research. Nevertheless, the conference paper correlations between citation rates and ChatGPT scores are much weaker (although not directly comparable) than the equivalent correlations for journal articles, so either or both might be much weaker indicators of research quality.

At the conference level, the correlations between citation rates and ChatGPT scores tend to be stronger than the paper level correlations, perhaps due to aggregation effects. They are also more variable between fields, presumably due to much smaller sample sizes.

The almost universally positive correlations between expert ranks and both geometric mean citation counts and mean ChatGPT scores within fields tend to support the hypothesis that both reflect conference average quality, but this is undermined by the expert ranks having an unknown amount of influence from conference citation rates. The results also suggest that citation rates are a slightly better indicator of conference quality overall in conference-based fields than are mean ChatGPT scores. This could be due to the influence of citation rates on conference rankings or may reflect that citation rates provide better evidence in this context.

At the field level, there was an unexpectedly high correlation between the average score for a field's conference papers and its average citation count. This is a strange relationship given that it is standard practice in scientometrics to compare citation counts only within fields or to field normalise them before comparing between fields on the understanding that there are technical reasons why citation rates vary between fields for reasons other than research quality (Moed, 2008). The field-level positive correlation found here suggests at least for possible hypotheses: (a) the standard scientometric assumption is false, (b) the situation is different for conference, (c) the technical reasons explain only some of the citation rate differences between fields, leaving a residual cross-field quality association between paper quality and uncorrected citation counts, or (d) ChatGPT quality scores are influenced by citation counts. It is not clear which is true at the moment.

## Conclusion

The results suggest that ChatGPT gives substantially different scores to citation rates, but could be better or worse as a source of research quality indicators for conferences in conference-based fields. This finding is not relevant to the majority of fields, for which conferences more rarely play an important information dissemination role. The evidence in this paper is consistent with both citation counts and ChatGPT scores being weak indicators of the quality of papers in conference-based fields, with indirect evidence suggesting that ChatGPT scores may be stronger evidence of research quality, but direct evidence from comparisons with expert conference rankings that citation rates are better than mean Chat GPT scores as indicators of the prestige of a conference.

Despite this finding, the weakness of the evidence suggests that extreme caution should be deployed if either is used to support a research evaluation. Conference papers may well have functions other than knowledge building, such as revealing the current state of the art in a technology field, which may be hard to capture with either indicator. Of course, since ChatGPT's estimates here are based on processing titles and abstracts rather than full text, they cannot be evaluations of the quality of papers but are guesses based on limited evidence. Thus, their utility derives solely from the extent to which they correlate positively with expert judgement.

Practical applications of the results here are the same as for citation-based indicators. These include supporting expert rankings of conferences (e.g., by providing a ChatGPT average score for each conference) and supporting evaluations of the quality of individual research papers (e.g., by providing a ChatGPT score for the paper and scores for comparable papers to norm reference against). In this role, ChatGPT has the advantage that its scores can be calculated immediately when a paper or conference proceedings is published: there is no need to wait several years for citation count data to mature.

Given the results, ChatGPT should not replace citation counts as a quality indicator for conference papers or conferences, but it would be reasonable to employ it for papers or

conference proceedings that have been published too recently (e.g., within the last one or two years) to have had time to attract citations.

**Acknowledgement**:
This study was funded by an ESRC grant UKRI1079.

# References

Aduku, K. J., Thelwall, M. and Kousha, K. (2017), "Do Mendeley reader counts reflect the scholarly impact of conference papers? An investigation of computer science and engineering", Scientometrics, Vol. 112, pp. 573-581.

Agar, J. (2012), Science in the twentieth century and beyond, Polity Press, Oxford, UK.

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.

Björk, B. C. and Solomon, D. (2013), "The publishing delay in scholarly peer-reviewed journals", Journal of Informetrics, Vol. 7 No. 4, pp. 914-923.

de Leon, F. L. L. and McQuillin, B. (2020), "The role of conferences on the pathway to academic impact: Evidence from a natural experiment", Journal of Human Resources, Vol. 55 No. 1, pp. 164-193.

Elsevier (2024), Content coverage guide, available at: https://www.elsevier.com/products/scopus/content (accessed 24 December 2024).

Freyne, J., Coyle, L., Smyth, B. and Cunningham, P. (2010), "Relative status of journal and conference publications in computer science", Communications of the ACM, Vol. 53 No. 11, pp. 124-132.

Greco, A. N. (2024), "The State of the Humanities and Scholarly Publishing to 2000", in Scholarly Publishing in the Humanities, 2000-2024: Marketing and Communications Challenges and Opportunities, Springer Nature Switzerland, pp. 17-36.

Haukoos, J. S., and Lewis, R. J. (2005). “Advanced statistics: bootstrapping confidence intervals for statistics with “difficult” distributions”, Academic Emergency Medicine, Vol. 12 No. 4, pp. 360-365.

Hauss, K. (2021), "What are the social and scientific benefits of participating at academic conferences? Insights from a survey among doctoral students and postdocs in Germany", Research Evaluation, Vol. 30 No. 1, pp. 1-12.

Henry, J. (2008), The scientific revolution and the origins of modern science, Bloomsbury Publishing.

Kang, D., Ammar, W., Dalvi, B., Van Zuylen, M., Kohlmeier, S., Hovy, E. and Schwartz, R. (2018), "A dataset of peer reviews (PeerRead): Collection, insights and NLP applications", in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1647-1661.

Kim, J. (2019), "Author-based analysis of conference versus journal publication in computer science", Journal of the Association for Information Science and Technology, Vol. 70 No. 1, pp. 71-82.

Kochetkov, D., Birukou, A. and Ermolayeva, A. (2021), "The importance of conference proceedings in research evaluation: A methodology for assessing conference impact", in International Conference on Distributed Computer and Communication Networks, Springer International Publishing, pp. 359-370.

Kocoń, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydło, D., Baran, J. and Kazienko, P. (2023), "ChatGPT: Jack of all trades, master of none", Information Fusion, Vol. 99, p. 101861.

Kousha, K. and Thelwall, M. (2024), "Factors associating with or predicting more cited or higher quality journal articles", Journal of the Association for Information Science and Technology (Annual Review of Information Science and Technology), Vol. 75 No. 3, pp. 215-244.

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.

Liang, W. et al. (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. et al. (2024b), "Can large language models provide useful feedback on research papers? A large-scale empirical analysis", NEJM AI, Vol. AIoa2400196.

Mabe, M. and Amin, M. (2001), "Growth dynamics of scholarly and scientific journals", Scientometrics, Vol. 51 No. 1, pp. 147-162.

Martín-Martín, A., Thelwall, M., Orduna-Malea, E. and Delgado López-Cózar, E. (2021), "Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: a multidisciplinary comparison of coverage via citations", *Scientometrics*, Vol. 126 No. 1, pp. 871-906.

Moed, H. F. (2006), Citation analysis in research evaluation, Vol. 9, Springer Science & Business Media.

Morus, I. R. (Ed.) (2022), The Oxford History of Science, Oxford University Press.

Oppenheim, C., Greenhalgh, C. and Rowland, F. (2000), "The future of scholarly journal publishing", Journal of Documentation, Vol. 56 No. 4, pp. 361-398.

Raby, C. L. and Madden, J. R. (2021), "Moving academic conferences online: Aids and barriers to delegate participation", Ecology and Evolution, Vol. 11 No. 8, pp. 3646-3655.

Recker, J. and Mendling, J. (2016), "The state of the art of business process management research as published in the BPM conference", Business & Information Systems Engineering, Vol. 58, pp. 55-72.

REF2021 (2019), Panel criteria and working methods, available at: https://2021.ref.ac.uk/publications-and-reports (accessed 24 December 2024).

Ruscio, J. (2008), "Constructing confidence intervals for Spearman's rank correlation with ordinal data: a simulation study comparing analytic and bootstrap methods", *Journal of Modern Applied Statistical Methods*, Vol. 7 No. 2, paper 7. https://doi.org/10.22237/jmasm/1225512360

Scherer, R. W., Meerpohl, J. J., Pfeifer, N., Schmucker, C., Schwarzer, G. and von Elm, E. (2018), "Full publication of results initially presented in abstracts", Cochrane Database of Systematic Reviews, No. 11.

Sugimoto, C. R. and Weingart, S. (2015), "The kaleidoscope of disciplinarity", Journal of Documentation, Vol. 71 No. 4, pp. 775-794.

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, Vol. 10 No. 1, pp. 1-19. doi:10.2478/jdis-2025-0011.

Thelwall, M. and Yaghi, A. (2025a), "In which fields can ChatGPT detect journal article quality? An evaluation of REF2021 results", Trends in Information Management, Vol. 13 No. 1, pp. 1-29.

Thelwall, M. and Yaghi, A. (2025b), "Evaluating the predictive capacity of ChatGPT for academic peer review outcomes across multiple platforms", Scientometrics, Vol 130. pp. 5285–5307. https://doi.org/10.1007/s11192-025-05287-1

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P. and Lowe, R. (2022), "Training language models to follow instructions with human feedback", Advances in Neural Information Processing Systems, Vol. 35, pp. 27730-27744.

Voorhees, E. M. and Harman, D. (1999), "The text REtrieval conference (TREC) history and plans for TREC-9", ACM SIGIR Forum, Vol. 33 No. 2, pp. 12-15.

Vrettas, G. and Sanderson, M. (2015), "Conferences versus journals in computer science", Journal of the Association for Information Science and Technology, Vol. 66 No. 12, pp. 2674-2684.

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

Wang, W., Bai, X., Xia, F., Bekele, T. M., Su, X. and Tolba, A. (2017), "From triadic closure to conference closure: The role of academic conferences in promoting scientific collaborations", Scientometrics, Vol. 113, pp. 177-193.