
Article:
Thelwall, M. orcid.org/0000-0001-6065-205X and Kurt, Z. (2025) Research evaluation with ChatGPT: is it age, country, length, or field biased? Scientometrics. ISSN 0138-9130

https://doi.org/10.1007/s11192-025-05393-0

Scientometrics
https://doi.org/10.1007/s11192-025-05393-0

Research evaluation with ChatGPT: is it age, country, length, or field biased?

Mike Thelwall[1] · **Zeyneb Kurt**[1]

Received: 9 October 2024 / Accepted: 22 July 2025

Abstract
Some research now suggests that ChatGPT can estimate the quality of journal articles from their titles and abstracts. This has created the possibility to use ChatGPT quality scores, perhaps alongside citation-based formulae, to support peer review for research evaluation. Nevertheless, ChatGPT's internal processes are effectively opaque, despite it writing a report to support its scores, and its biases are unknown. This article investigates whether publication date and field are biasing factors. Based on submitting a monodisciplinary journal-balanced set of 117,650 articles from 26 fields published in the years 2003, 2008, 2013, 2018 and 2023 to ChatGPT 4o-mini, the results show that average scores increased over time, and this was not due to author nationality or title and abstract length changes. The results also varied substantially between fields, and first author countries. In addition, articles with longer abstracts tended to receive higher scores, mostly due to such articles tending to be better (e.g., more likely to be in higher impact journals) but also partly due to ChatGPT analysing more text. For the most accurate research quality evaluation results from ChatGPT, it is important to normalise ChatGPT scores for field and year and check for anomalies caused by sets of articles with short abstracts.

Keywords ChatGPT · Research impact · Publication date · Research excellence framework

Introduction

Expert review of the quality of published research outputs essential for appointments, promotion, and tenure. Post publication expert review also occurs systematically in some national research evaluation exercises (e.g., Technopolis, 2024). The use of bibliometrics to support, but not replace, expert review can improve score accuracy when experts are uncertain and might be used to reduce the burden on reviewers (e.g., by using two instead of three) when the bibliometrics are known to be accurate enough. This use is consistent with the wider considerations of the Leiden Manifesto (Hicks et al., 2015), the Metric Tide (Wilsdon et al., 2015), and the Coalition for Advancing Research Assessment (coara.eu). Improving the value of bibliometric indicators, the focus of much scientometric research,

✉ Mike Thelwall
m.a.thelwall@sheffield.ac.uk

[1] Information School, University of Sheffield, Sheffield, UK

or finding an improved alternative, as in the current paper, therefore has the potential to save expert time and/or increase the accuracy of their decisions.

Artificial Intelligence (AI) methods have been proposed for various aspects of evaluation systems, such as reviewer selection (Holm et al., 2022), with the almost complete exception of the evaluation task (e.g., Carbonell Cortés et al., 2024), where either peer review or bibliometric-supported peer review dominate. Ethical and legal considerations are important here, including for processing any private data and ensuring that evaluators are aware of the strengths and weaknesses of any AI score predictions that they are shown (Carbonell Cortés et al., 2024). At the moment, there do not seem to have been any proposals to completely replace expert evaluations for research in any context. These would raise substantial issues, including for bias and systemic effects (Rushforth & Hammarfelt, 2023; Ye et al., 2024).

Previous research has found that ChatGPT might provide an alternative to citation-based indicators in support of human expert review, or for some theoretical investigations of science that currently use citation analysis (e.g., Chen et al., 2015). This is because, when fed with expert review instructions, ChatGPT can provide quality scores for journal articles that positively correlate with human expert scores in all fields of scholarship except perhaps clinical medicine (Saad et al., 2024; Thelwall, 2024, 2025; Thelwall & Yaghi, 2024). Moreover, ChatGPT's scores have a higher correlation with human expert scores than do citation-based indicators or traditional AI in most fields (Thelwall & Yaghi, 2024). Despite these promising statistical results, and the fact that AI evaluations of academic proposals are now supporting at least one funding agency (Carbonell Cortés et al., 2024), nothing is known about any biases in ChatGPT results. In other contexts, the potential for AI systems to replicate human prejudices and even invent new partialities has been shown (Ntoutsi et al., 2020). It would therefore be unwise to assume that ChatGPT's research quality evaluations are exempt or to use it for important evaluations without prior investigations into potential biases.

In response to the above concern, this article primarily focuses on one important potential bias, publication year. When human experts assess articles for research quality, they can be expected to consider the publication date when evaluating novelty and significance because, for example, earlier articles on a topic would tend to be more important than later articles, other factors being equal. They might be judged more original by covering an under-researched or new topic or more influential by making first investigations that later contributions then refined. This article therefore assesses whether ChatGPT's average quality scores vary based on the publication years of articles, which ChatGPT is typically not told. The main concern is that it would tend to give lower scores to older articles because it does not consider their value relative to when they were published. Field differences in citation rates are well known in citation analysis (Dunaiski et al., 2019) and so this is another source of potential bias that should be investigated. Finally, since optimal ChatGPT quality scores are obtained from the title and abstract of an article, ignoring its full text, it is logical to also assess whether the combined length of these associates in any way with ChatGPT scores. The main research questions are therefore as follows.

- RQ1: Do average ChatGPT scores vary systematically by year?
- RQ2: Do average ChatGPT scores vary between fields?
- RQ3: Do average ChatGPT scores associate with title and abstract length?

Whilst country differences may not be due to biases but instead a result of national factors, such as the degree of financial support for research, it is important to identify any

such differences for text-based indicators because of the possibility that countries are disadvantaged if they have few native English speakers. If English speaking countries *consistently* outscored others with ChatGPT, then this would suggest at least the possibility of a language bias, although the results would need to be interpreted in the context of national research support differences.

- RQ4: Do average ChatGPT scores vary between countries?

This article also reports whether ChatGPT scores associate with citation counts because the biases of citation-based indicators have been extensively investigated (e.g., Didegah, 2014) and this evidence provides additional context to the main results.

- RQ5: Do ChatGPT scores correlate with citation counts?

Bias in text processing AI systems

Extensive research has shown that AI systems can be biased through design choices, by learning biases from their training data, or as an unintended side effect of their design and training data (e.g., Caliskan et al., 2017). As a simple example, an AI system designed to select suitable candidates for jobs based on their CVs might learn that males are more likely to be interested in male-associated jobs like carpentry and then use gender as an indicator for job preferences (e.g., via first names), avoiding selecting females for them (Wilson & Caliskan, 2024). This example illustrates the potential for bias amplification (Wang et al., 2024). More generally (and an issue in the current paper), many AI systems tend to be trained, evaluated or refined on data primarily of Western origins, resulting in many potential biases against global majority cultural considerations (Peters & Carman, 2024).

Bias in large language models (LLMs)

Like other machine learning systems, LLMs can learn biases from the texts that they have ingested. Here the term bias covers both disparities (e.g., usually choosing male pronouns for doctors of an unknown gender) and cognitive biases in the sense of making sub-optimal decisions due to stereotypes or associations overriding the evidence. Press coverage of apparent political, gender or other biases in LLMs has led to efforts to reduce their biases. Now, major LLMs are unlikely to agree with stereotypes asked about by users (e.g., "Are women poor racing drivers?"; ChatGPT 4o: "Not at all. []"). This may reflect successful debiasing training strategies (Li et al., 2024). Nevertheless, older LLMs that are still in use, such as versions of Meta Llama 3.1, exhibit clear biases (Kumar et al., 2025). Moreover, whilst some newer LLMs can now pass formal bias tests from standard benchmarks that mainly test for explicit bias (Lin, & Li, 2025) they still seem to have biases in the sense of associating some words with categories such as gender (ChatGPT 4o in: Bai et al., 2025; Kotek et al., 2023). Moreover, it is possible that the more complex architectures of larger LLMs allows them to both avoid obvious bias and more strongly learn implicit biases through association (Zhao et al., 2025). For example, the largest of the related Llama 3.1 models in one test had the strongest association bias (Kumar et al., 2024). This suggests

that the largest modern LLMs have outputs that are influenced by stereotypes, but not in an obvious way.

To illustrate the potential influence of learned biases, major LLMs seem to have learned stereotypes about ethnic minority populations in the USA that influence recommendations for medical care (Omiye et al., 2023). Bias can occur even when medical guidelines are clear and have been ingested by a LLM because its non-expert training data also influences its output (e.g., Bender et al., 2021; Gururangan et al., 2020). The fundamental problem here is perhaps that the LLM does not fully distinguish between accurate information and low-quality information in its inputs.

Prompt-based strategies to mitigate against LLM bias

End users may wish to identify and mitigate against biases in LLM outputs. It has been suggested that LLMs may give misleading explanations for biased results to rationalise them (Kotek et al., 2023), making it more difficult to detect bias within individual outputs. Thus, asking an LLM if its decision was biased is not an effective strategy.

Different prompt-based strategies have been tested to mitigate bias. These include (a) warning against cognitive bias, (b) giving a few examples of correct or incorrect behaviour with respect to a single identified potential bias (a fewshot learning strategy variant), and (c) asking the LLM to rewrite the original prompt to reduce the chance that it will elicit cognitively biased answers. Of these three, warnings produce small improvements for some context and rewriting helps with the largest LLMs when the original prompts are problematic (Echterhoff et al., 2024).

Methods

The research design was to take samples of research articles from a range of fields and a range of widely spaced years, controlling for as many variables as possible, then compare the average scores by year and field for an initial descriptive analysis and then use regression to compare the effects of all variables simultaneously. A simultaneous comparison of all variables through regression guards against one bias being a second order effect of another.

Although previous studies have shown that more accurate scores can be obtained by ChatGPT if articles are submitted multiple times (Thelwall, 2024, 2025), accuracy is not of primary interest for the current study (except RQ5, which is provided for context) and so each article was submitted only once.

Data

A dataset was needed of articles published in different years but with the same average quality each year. Unfortunately, it is impossible to be sure that the average quality of two sets of papers published at different points in time is the same because all aspects of academic scholarship evolve. The simplest approach is to take a random sample from the subject categories of a scholarly database, such as the Web of Science (WoS) or Scopus, that are quality controlled. This is not ideal, however, because the coverage of these databases can shift substantially when a new tranche of journals is added (Moed et al., 2018). The situation is exacerbated if indexing policy changes occur, such as to add journals from an emerging research country. A better, but still imperfect, approach might be to select one

or a small set of core journals from each field to take samples from, on the basis that the quality of journals seems to be relatively stable. This is unproven and journals can change in nature when a new editor takes over or perhaps after a deliberate shift in scope (Martin, 2016; Wilhite et al., 2019) but there does not seem to be a more reliable approach.

Based on the above discussion, a journal-based sampling method was designed to minimise likely biases. Scopus was chosen as the database to select the data from, although WoS would have been equally appropriate. The years 2003, 2008, 2013, 2018 and 2023 were selected with five-year gaps to identify overall trends over the past two decades. A relatively long period was needed in case changes between adjacent years were too small to be detected statistically. The Scopus records were downloaded in January–March 2024. The 26 non-general categories in Scopus were used as the fields to analyse. Field definitions are controversial (e.g., Waltman & Van Eck, 2012) and fields evolve over time (Porter & Rafols, 2009), but the Scopus definitions are at least transparent in the sense that the underlying sets of journals are openly published (www.elsevier.com/en-gb/products/scopus/content). Each field was represented from the set journals that had articles published in each of the five years. Journals were selected when they were not also classified in other Scopus categories to avoid overlaps between fields. Although some overlap would be reasonable, this rule has the advantage of excluding huge multidisciplinary journals that would have many articles that are out of scope for some of the fields in which they are classified.

Article titles and abstracts were extracted from Scopus and automatically cleaned of copyright statements using Webometric Analyst (github.com/MikeThelwall/Webometric_Analyst) for submission to ChatGPT. Some articles in Scopus are short form, letters or corrections and are not flagged as such in the database. These typically seem to have a short abstract. For this reason, and because some articles did not have an abstract in Scopus, a minimum abstract length threshold of 786 characters, after removing abstract copyright statements, was set for articles to be selected. For the journals selected, this represented ignoring 25% of Scopus articles with short or no abstracts. This threshold was chosen as a round number (in percentage terms) and as, heuristically, long enough to judge the content of a full article. For example, a threshold of 10% equates to 403 characters which is short, and 20% equates to 682 characters, which still seemed too short for a meaningful description of an academic article.

Articles were selected at random from all journals matching these criteria, for a total of 1000 for each year. The random selection process was conducted by journal, so each journal had the same number of articles in each year. Because of this, the overall number of articles per broad field varied slightly from the target 1000.

In summary, for each of the years 2003, 2008, 2013, 2018 and 2023 and each of the 26 non-general Scopus categories, up to approximately 1000 articles were selected with a random number generator (without replacement) from the set of all articles with substantial abstracts in journals that were exclusive to that category and that had published in all five years (Fig. 1). This gave a total of just under $1000 \times 5 \times 26$ articles to submit to ChatGPT.

ChatGPT procedure

The ChatGPT setup was the same as for a previous analysis of journal articles submitted to the UK REF (Thelwall & Yaghi, 2024). In summary, ChatGPT was supplied with the appropriate evaluation guidelines given to REF assessors. These guidelines were used because they have been previously tested with ChatGPT and are reasonably detailed guidelines leading to specific scores (1*, 2*, 3* or 4*). There are four guidelines, and the most relevant one was used for each field. In particular, the health and life sciences

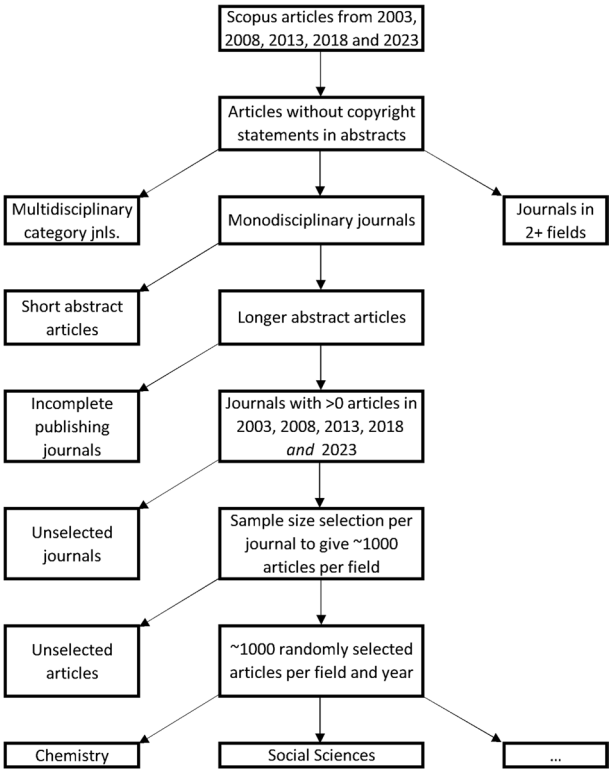

Fig. 1 The main sample selection stages

guidelines (REF Panel A) were applied to Scopus fields 11, 13, 24, 27, 28, 29, 30, 32, 34, 35, 36; the physical sciences and engineering guidelines (REF Panel B) were applied to fields 15, 16, 17, 18, 19, 21, 22, 23, 25, 26, 31; the social sciences guidelines (REF Panel C) were applied to 14, 20, 33; and the arts and humanities guidelines (Main Panel D) were applied to 12. These numbers can be converted to fields by adding 00 and consulting https://service.elsevier.com/app/answers/detail/a_id/15181/supporthub/scopus/.

Each article was submitted to ChatGPT 4o-mini using the ChatGPT API, with a separate chat session for each article. This interface does not use the submitted material for model building (openai.com/consumer-privacy), which avoids any potential feedback within the experiment as well as any copyright concerns. The scores were extracted by a program designed to apply a set of pattern-matching rules to identify scores from ChatGPT reports (github.com/MikeThelwall/Webometric_Analyst, AI menu). When its rules failed, the first author was shown the ChatGPT output and prompted for a number. The scores were usually whole numbers, but occasionally separate scores were provided for the three individual aspects of research quality considered in the REF (originality, rigour, significance) and these where then averaged for the overall score. This final ChatGPT score for an article is thus a number between 1 (nationally relevant) and 4 (world leading) that is ChatGPT's evaluation of the research quality of the article, using the UK REF research quality criteria.

Descriptive analysis

Average ChatGPT scores were calculated for each year/field combination. The results were then graphed for each field to identify any changes over time and between fields. Pearson correlations for ChatGPT score and the log of the abstract length were then calculated for each year and field combination to identify any large scale underlying trends.

Regression

Ordinary least squares regression was used to check whether publication year could be a second order effect of changes over time in abstract length or author affiliation country, and to check for the importance of the latter variables. For this, the ChatGPT score was used as the dependent variable, with the following independent variables: publication year; title and abstract length combined (number of characters); first author country (ten binary-coded variables). First author country was recorded for the ten countries with the most Scopus publications. Although it would have been possible to include a separate binary variable for each country in the world, this would lead to overfitting so a focus on the top ten publishing countries seems adequate to capture the main national influences. For additional regressions, article field and year normalised citations were also used as a dependent variable, with the Normalised Log-transformed Citation Score (NLCS) formula (Thelwall, 2017) and journal citation rates with the journal NLCS—the mean NLCS for all articles in the journal (JNLCS).

Both abstract and title length and the log of abstract and title length were tried as independent variables. They had high Variance Inflation Factors (VIFs) between them (typically above 20) so only one could be retained due to the multicollinearity. The log version was used because it was statistically significant in more regressions combining both. After this decision, VIFs were calculated for all variables, with most values being close to 1 and none above 2, so the multicollinearity is not strong enough to correct for. Q-Q plots were broadly consistent with normally distributed residuals.

Results

RQ1-3: descriptive analysis

The descriptive analysis of average ChatGPT scores shows differences between fields and systematic differences between years. For all 26 broad fields, the ChatGPT average for 2023 was higher than for 2003, and in 101 out of 104 cases, the ChatGPT average for each year (2008, 2013, 2018, 2023) was higher than the average from the previous year. There was also a positive relationship between publication year and ChatGPT score at the article level, with Pearson correlation coefficients ranging between 0.077 (Mathematics) and 0.306 (Business, Management and Accounting), with a mean of 0.181. From the perspective of the article-level influence of publication year, the R^2 values corresponding to these correlation coefficients (not the linear regressions, which are discussed in the next section) show that the amount of variance in the ChatGPT scores

explained by the publication year varied between 0.6% and 9.4%, with an average of 3.6%. This indicates a relatively small influence.

There are also substantial differences in average ChatGPT scores between fields (Fig. 2). For example, the averages for all five years of the top three fields (Biochemistry, Neuroscience, Mathematics) are higher than the averages of all fields in all years. Conversely, the averages for all five years of the lowest field, Veterinary, are lower than the averages for all years of all fields except the bottom four fields.

Articles with longer abstracts tend to have moderately higher ChatGPT scores in some fields but not others, with a slight tendency for weaker correlations in more recent years (Fig. 3). From the perspective of the article-level influence or association with abstract length, the R^2 values corresponding to these correlation coefficients (again, not the linear regressions) reveal that the amount of variance in the ChatGPT scores explained by abstract log-length variable varied between 0.0% and 11.7%, with an average of 3.3%. This indicates a smaller influence or association than for publication year overall, but larger in some fields.

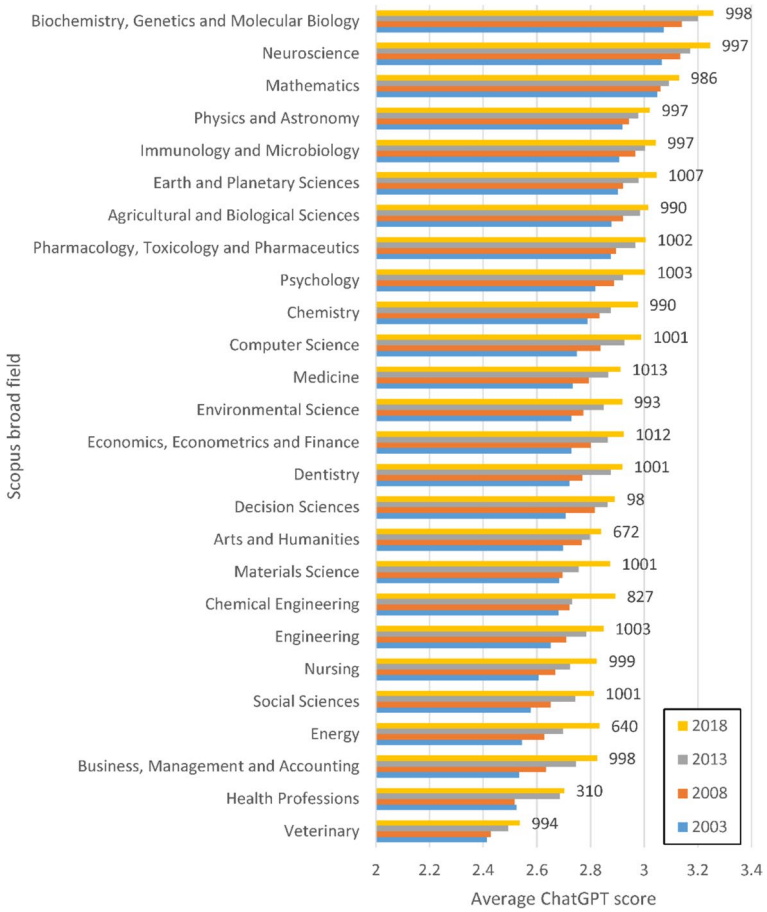

Fig. 2 Average ChatGPT scores for articles from journals exclusively in each of 26 Scopus broad fields, by publication year. Fields are annotated with their sample sizes (number of articles in each of the 5 years)

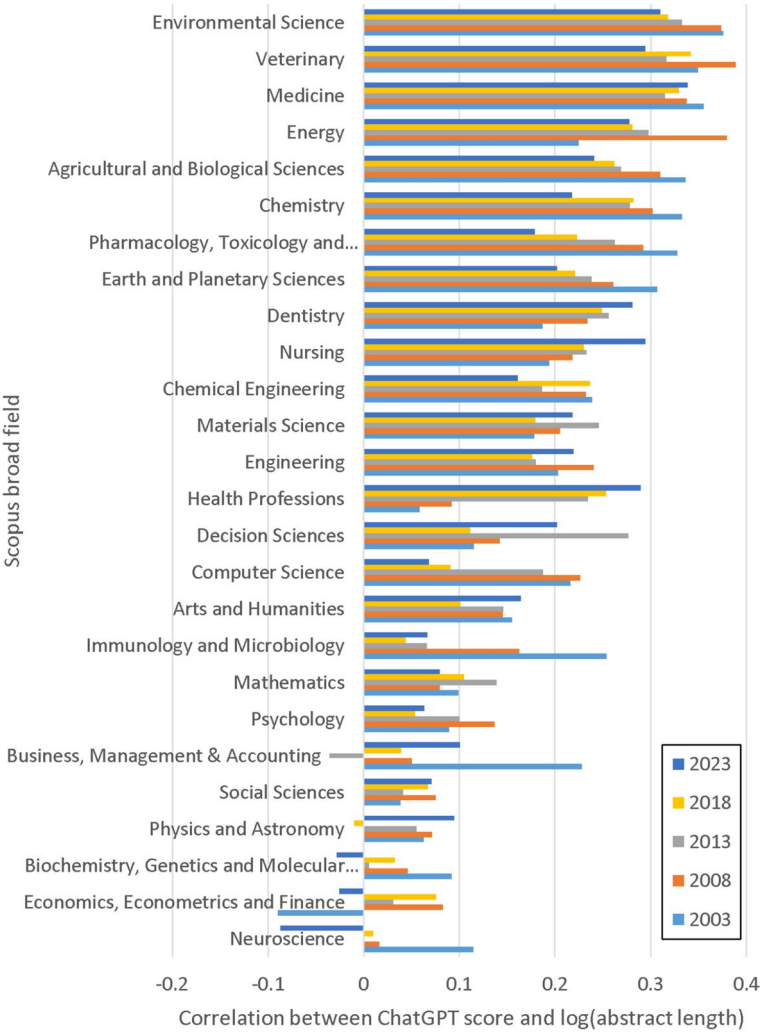

Fig. 3 Pearson correlations between ChatGPT score and the log of the abstract length by year and field. Sample sizes are as in Fig. 2

RQ1-4: regression results

This section reports the aggregate results from the 26 separate field-based regressions. By taking all variables into account simultaneously, they help to exclude the possibility that any association is a second-order effect of one of the other independent variables.

In all 26 regressions, the year variable coefficient was positive and statistically significant, suggesting that the reason for the lower average ChatGPT scores for older articles is not due to changes in either the typical countries of the first authors or abstract length (Fig. 4). Nevertheless, articles with longer abstracts tend to get higher scores in all 26 fields, and the coefficient is statistically significant in 23 of these fields. This is despite

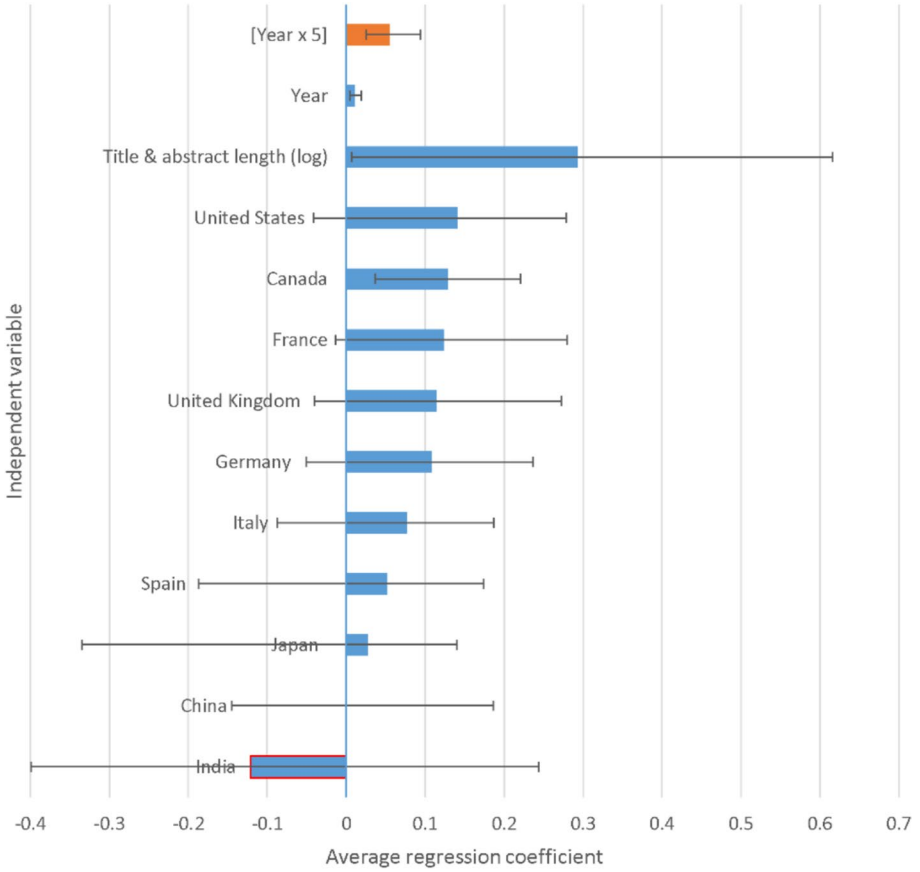

Fig. 4 Summary of the coefficients extracted from 26 regressions for ChatGPT scores against the variables above. [Year×5] is a dummy variable for article publication year, multiplied by 5 to illustrate the influence of a five year difference in publication years. Country names are for first authors, and only the ten countries with the most articles in Scopus were included. The coefficients can be compared between countries but none of the other coefficients should be compared. Dependent variable: ChatGPT score (1–4). Bars represent average regression coefficients across 26 field-specific OLS models; error bars indicate the minimum and maximum values observed across those models. Table 1 (column numbered 4) reports the number of statistically significant coefficients for each dependent variable

the minimum length abstract cutoff (see methods) that was designed to exclude short form contributions.

Although Canadian first authors were the only national set found to associate with higher ChatGPT scores in all fields (statistically significant in 21), most of the other countries with the most journal articles in Scopus also tended to be associated with higher ChatGPT scores. The main exception, India, has a low per capita research investment (Gross Expenditure on R&D as a percentage of GDP: https://data.uis.unesco.org/index.aspx?queryid=74), so its researchers are less supported. Considering the coefficients of Italy, France, and Germany, there is not a clear pattern of English-speaking nations having an advantage over others. Moreover, given that the REF criteria are important for UK academics, it is perhaps surprising that its authors do not dominate for the ChatGPT score derived from the same criteria.

Table 1 The number of Scopus broad fields (out of 26) for which the dependent variable coefficient is statistically significant ($p < 0.05$) in the regressions underlying the figures

Dependent variable	Figure number				
	4	6	7	8	9
Jnlcs	–	26	–	26	–
Nlcs	–	–	–	–	26
Year	26	26	21	7	26
Title and abstract length (log)	23	25	18	17	23
United States	24	24	19	4	24
Canada	21	18	11	4	20
France	18	15	10	3	17
France	18	15	10	3	17
United Kingdom	22	19	18	6	20
Germany	18	21	15	1	20
Spain	11	5	10	1	10
Japan	13	10	14	9	11
China	18	11	14	11	15
India	17	15	12	8	16

Variations between coefficient strengths between fields can be examined with online graphs of each one (https://doi.org/10.6084/m9.figshare.28328852). This shows, for example, that the Health Professions field has the largest negative coefficient for Italy, Japan, and the USA.

The R^2 values of the models varied from 0.05 to 0.21, with a mean across fields of 0.12. Thus, on average, the factors considered above only account for 12% of the variation in the ChatGPT scores, so they do not have an overriding importance within a field. These values will be slight overestimates because cross-validation was not used to allow accuracy to be checked on different datasets to those used for model building.

RQ5: correlation with citation counts

For all fields, citation counts correlated positively with ChatGPT scores for the articles examined, although with substantial differences between fields (Fig. 5). Correlation differences between years may be due to randomness inherent in the sample selection rather than due to systematic differences since the average of the 26 correlations for each year are similar without a monotonic increasing or decreasing trend (2003: 0.247; 2008: 0.232; 2013: 0.242; 2018: 0.233).

Discussion

This article is limited by the sample selection method, which excludes articles in multidisciplinary journals and articles not indexed by Scopus. These may follow a different trend. The results may also be different for other LLM-based systems and may improve over time as each system evolves. The ChatGPT scores reflect the UK's national research quality definition but there are many others (Langfeldt et al., 2020), with Global South versions likely to be substantially different (Barrere, 2020). Moreover, if

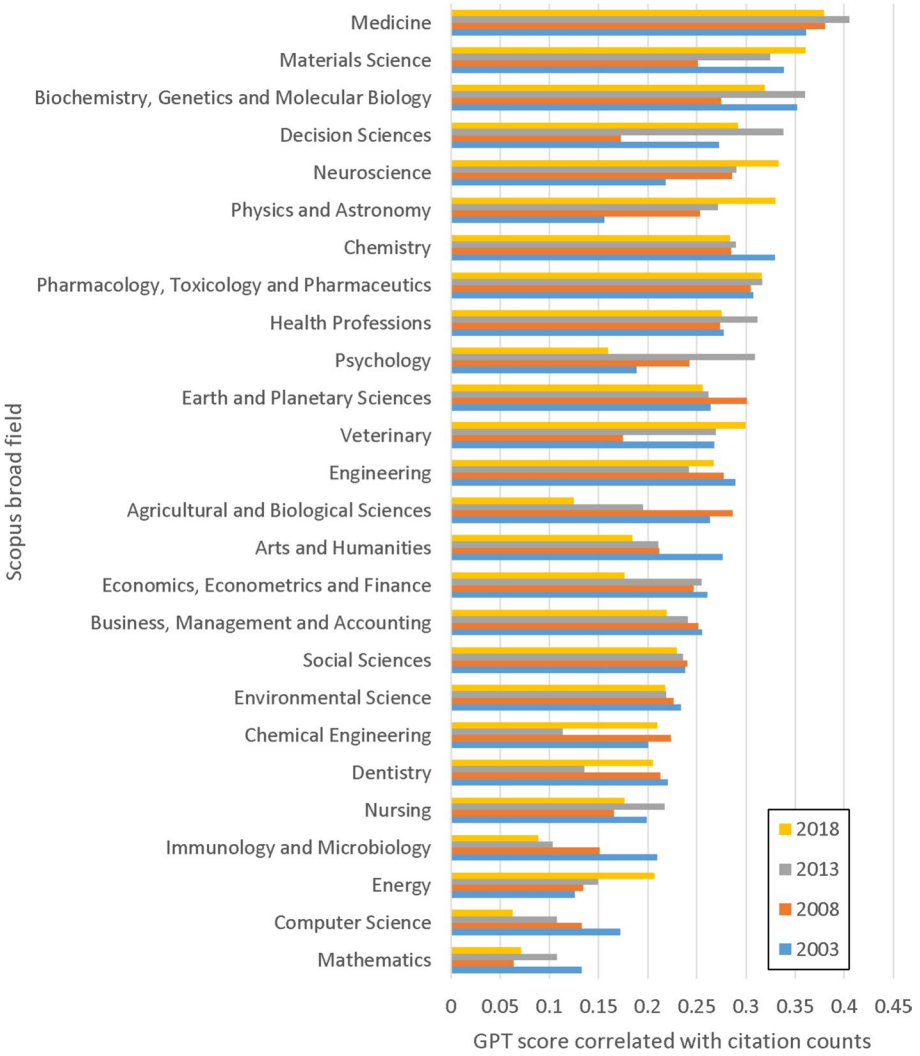

Fig. 5 Spearman correlations between ChatGPT scores and Scopus citation counts. 2023 is excluded due to insufficient time for citations to accrue (Wang, 2013)

ChatGPT scores start to be widely used in research evaluation then authors may try to game them with their abstract designs, which may undermine the approach.

A more subtle limitation is the assumption that the differences between years, fields, and ages constitute a bias rather than reflecting existing disparities. For example, if there was an objective reason to believe that older research was worse than newer research or that some fields produced better studies then the disparities found here would not be biases in the sense of unfair differences (Traag & Waltman, 2022). Nevertheless, in the absence of an objective ground truth for research quality and given substantial disciplinary differences in fine grained evaluation criteria, it seems likely that research managers would expect research evaluation scores to be approximately the same for different

years, and fields. Finally, the focus of this paper is on biases in ChatGPT scores rather than on other issues related to implementing an LLM-based research evaluation system.

It is not known whether there is a different ChatGPT prompt that would have elicited less biased responses. For example, the prompt could be modified to include the age and/or field of the article with a request to score against the norms of the field and year. Averaging multiple ChatGPT scores rather than single scores seems unlikely to reduce bias, although it would improve accuracy. For example, the current data suggests that older articles tend to get lower ChatGPT scores and the same would be true for ChatGPT scores based on averaging five scores per article rather than individual scores. It would also be useful to get a more transparent evaluation from ChatGPT that would illuminate the source of disparities. Its reports did not mention author nationality, publication year or abstract length, but asking explicitly about these factors in prompts might generate new insights. Preliminary experiments with this did not provide any useful information, however. Of the previously suggested anti-bias strategies (Echterhoff et al., 2024), adding an explicit admonition to avoid bias might be worth trying.

RQ1: The tendency for ChatGPT research quality scores to be higher for more recent articles has not been previously tested but the result is expected given that originality is a component of REF quality and is clearly time dependent in the sense that ChatGPT might consider a just-published paper to be more original than a decade old paper, given its knowledge of the world. Moreover, research quality might well have improved over time, on average in some or all fields. This is perhaps clearest in science fields where knowledge is cumulative and technical fields where technology improves. For example, this year's AI paper might tend to give substantially better results than decades old papers due to LLMs and increased computing power. The converse relationship exists for recent citation counts, which accrue over time. Very old papers may get fewer citations, however, since articles mostly cite recent work and there were shorter reference lists in the past (Larivière et al., 2008). From a research evaluation perspective, it seems undesirable to allow older research to be given lower scores because that would penalise researchers for not achieving levels that would have been more difficult at the time of publication. Thus, normalisation for publication year seems appropriate.

RQ2: Field differences in average ChatGPT scores have not been found before but were to be expected given very different field norms in publishing styles (Puuska, 2014). From a research evaluation perspective, given the apparent impossibility of norm referencing research quality between fields, expecting similar quality scores for all fields seems to reasonable, thus requiring field normalisation.

RQ3: In theory, the tendency for articles with longer abstracts to get higher ChatGPT scores could be a direct effect of better articles needing longer abstracts, a side effect of better journals allowing longer abstracts, or the inclusion of short form contributions with shorter abstracts in some or all fields in the dataset. Repeating the regressions with average journal citation rate as an additional independent variable substantially reduces the explanatory strength of title and abstract lengths (Fig. 6). This suggests that most, but not all, of the association between higher ChatGPT scores and longer abstracts is due to more cited journals tending to have longer abstracts.

An examination of the five fields with the highest correlations in Fig. 3 also suggests that the second and third options play a role as follows.

- **Better journals may allow longer abstracts** Journals publishing longer articles tend to attract higher ChatGPT scores in 2023 Environmental Science (average journal level Pearson correlation between article log-length and ChatGPT score: 0.451), Veterinary

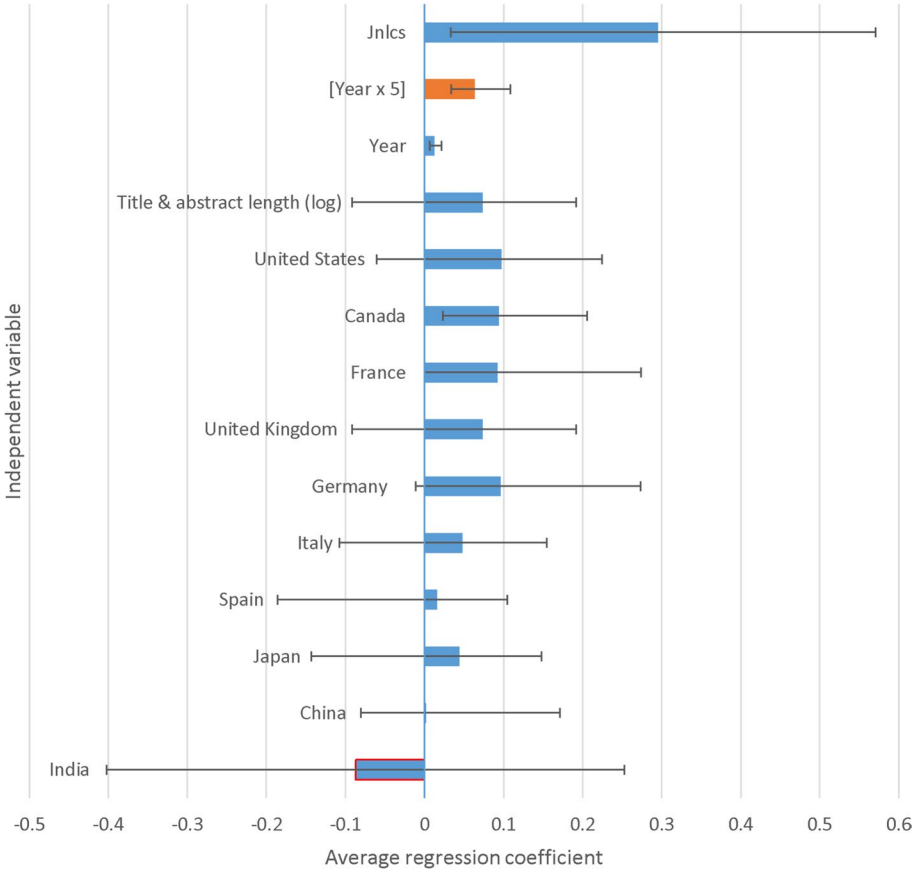

Fig. 6 As Fig. 4 except with journal citation rate (Journal average NLCS) as an additional independent variable. Dependent variable: ChatGPT score (1–4). Bars represent average regression coefficients across 26 field-specific OLS models; error bars indicate the minimum and maximum values observed across those models. Table 1 (column numbered 6) reports the number of statistically significant coefficients for each dependent variable

(0.331), Medicine (0.363), Energy (0.434), and Agricultural and Biological Sciences (0.278). For Veterinary, Medicine, and Agricultural and Biological Sciences, journals from countries where English is not the spoken language were overrepresented in the set of journals with the lowest scores. In some of these, the articles were primarily not in English so the abstracts and titles were translated or constructed separately, either of which may have had a shortening effect. These journals also tended to publish less cited work and some of their articles were summaries of a topic for a local audience rather than primary research (e.g., "Dental diseases in pet rabbits—anatomy, etiology, clinical presentation, differential diagnoses and diagnostic imaging" in *Kleintierpraxis*, a journal "For veterinarians in clinics and practices"). These seemed to have shorter abstracts that described the topic but did not need to mention methods. In Environmental Science, a set of apparently prestigious journals (e.g., high journal impact factors) had long abstracts and high average ChatGPT scores, suggesting that the ChatGPT score association with abstract length might be at least partly a second order effect of

journal quality rather than bias. Differences between journals seems to be more important than differences within journals, although the latter is also a factor. This is evident from the fact that correlations between abstract length and ChatGPT score tended to be positive, albeit low: sample size weighted averages of the within-journal correlations for 2023 are Environmental Science (0.169), Veterinary (0.248), Medicine (0.195), Energy (0.108), and Agricultural and Biological Sciences (0.131).

- **Short form contributions with shorter abstracts may tend to get lower scores** This was evident for Veterinary, where some journals had Brief Reports and Case Reports in addition to Standard/Research articles. Case Reports were also in the Medicine dataset. It seems reasonable to expect that short form articles tend to be lower quality than standard articles, so this does not suggest a ChatGPT length bias.

Overall, the association with abstract length is therefore primarily a non-bias second order association, at least from the perspective of REF quality definitions. Nevertheless, there is also a small bias in terms of ChatGPT favouring longer abstracts.

RQ4: International differences in ChatGPT scores have also not been investigated before although international differences in citation counts are understood and used to inform or monitor science policy (Gov.UK, 2022). The use of citations for international comparisons is problematic, however, because of a tendency for national self-citations and the extent to which researchers publish in national literature that is not indexed in major citation databases (Adams, 1998). ChatGPT does not have the same problems, and the current results suggest that English-speaking countries may not have an advantage, but the evidence is thin and there may be other sources of bias, such as towards research that helps richer countries. The lack of clearer evidence of international biases is perhaps surprising in the context of prior evidence that the largest LLMs have association biases and could easily have picked up cognitive/association biases (Bai et al., 2025; Kotek et al., 2023), such as that US-focused research tends to be more prestigious in some fields. LLMs may also have associated more academic writing styles from more fluent English speakers with higher scores, but there was insufficient evidence to show this. Speculatively, these potential associations may have been too weak to have a big influence given the complexity of the task. It is not clear whether national differences should be normalised for in ChatGPT scores for research evaluations.

RQ5: The positive correlations with citation counts in all fields found above echo the positive correlations between field normalised citation counts and the article-level expert quality scores of REF2021 (Thelwall et al., 2023). This is consistent with, but does not prove, the hypothesis that ChatGPT scores tend to positively associate with article quality. This association has been found directly for two small samples (Saad et al., 2024; Thelwall, 2024, 2025) and for an indirect quality measure (Thelwall & Yaghi, 2024), so this adds to a growing body of evidence about the relationship. This study used single ChatGPT scores whereas previous research has found higher correlations (with quality scores) when averaging thirty ChatGPT scores per article (Thelwall & Yaghi, 2024), so the underlying correlations might be double those in Fig. 5. The positive Medicine correlation is unexpected given the negative correlation between REF departmental scores and ChatGPT scores for the UK REF Clinical Medicine category previously found (Thelwall & Yaghi, 2024). This may indicate substantially different scopes for the two categories, a departmental selection effect in the previous study that selected articles from a few departments, or a UK anomaly, but probably not an underlying difference between citations and research quality in this field, given the known positive correlation between these two variables (Thelwall et al., 2023).

In terms of biases, the difference between ChatGPT scores and article citation rates (NLCS) can be compared by repeating the original regressions (Fig. 4) after replacing ChatGPT Scores with article citation rates (NLCS) as the dependent variable (Fig. 7). The overall pattern is similar, except that the influence of publication year is smaller (it should be zero by NLCS design, suggesting a second order interaction with other variables), abstract lengths are less influential, national differences vary more between fields (wider ranges, as reflected in the error bars). There are also some individual country differences, such as a larger NLCS bias towards Italy and a bias against Japan.

If the journal citation rate (average journal NLCS) is added as an additional independent variable, then this reduces the predictive power of the other variables for the NLCS dependent variable. Thus, the publishing journal overrides the authors' countries to some extent (recall that the bar sizes are not directly comparable due to different measurement units) (Fig. 8).

Finally, GPT scores can be compared directly against article citation rates by adding NLCS as an additional independent variable to the original regression (Fig. 4) to give

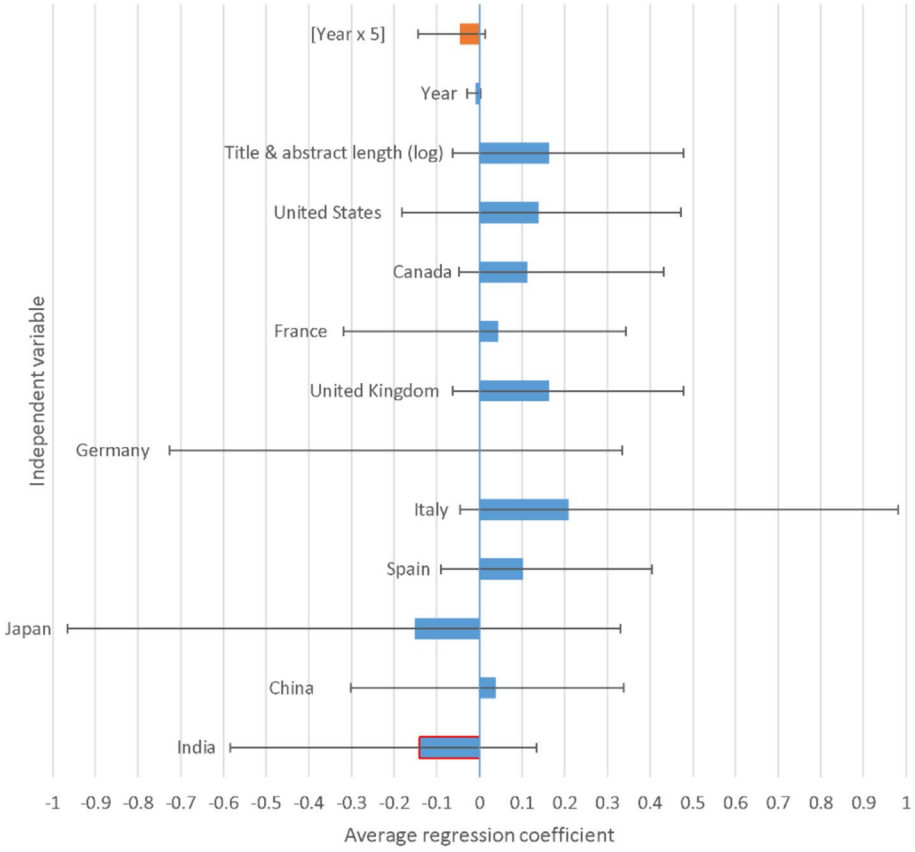

Fig. 7 As Fig. 4 except with field and year normalised article citation rate (NLCS) as the dependent variable. Dependent variable: Article NLCS citation rate. Bars represent average regression coefficients across 26 field-specific OLS models; error bars indicate the minimum and maximum values observed across those models. Table 1 (column numbered 7) reports the number of statistically significant coefficients for each dependent variable

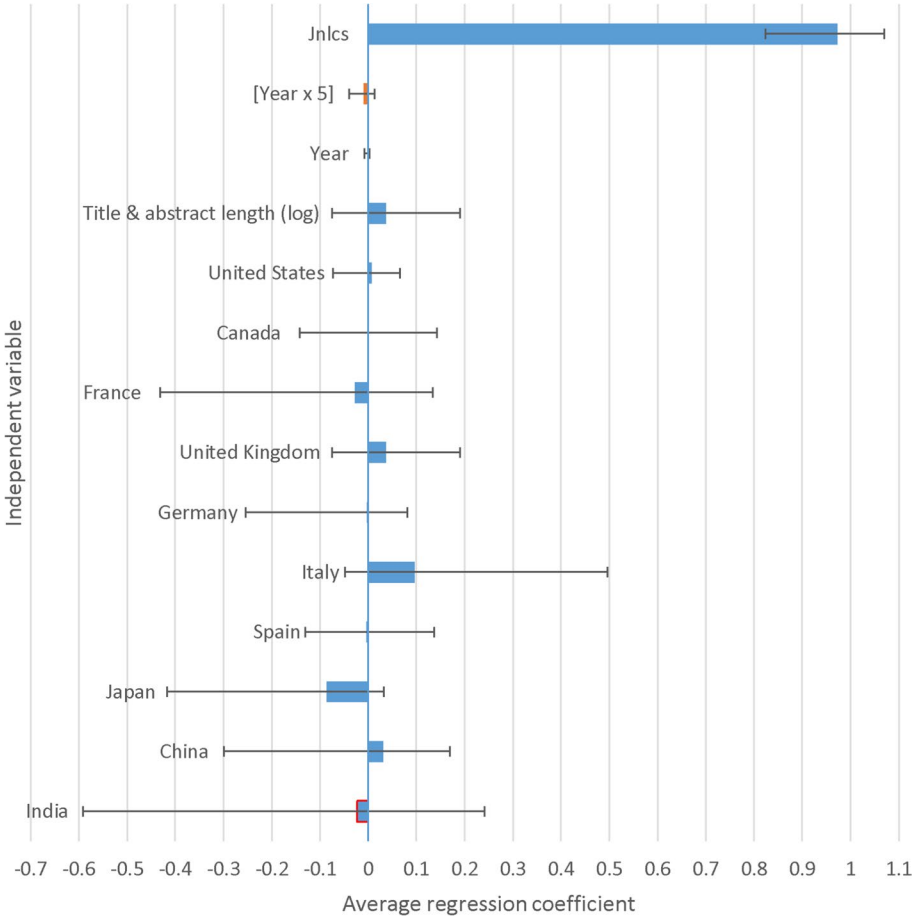

Fig. 8 As Fig. 7 except with journal citation rate (Journal average NLCS) as an additional independent variable. Dependent variable: Article NLCS citation rate. Bars represent average regression coefficients across 26 field-specific OLS models; error bars indicate the minimum and maximum values observed across those models. Table 1 (column numbered 4) reports the number of statistically significant coefficients for each dependent variable

regressions that consider article citation rates (Fig. 9). This tends to confirm the patterns in the original regression, although reducing the importance of title/abstract lengths.

Conclusions

The results give strong evidence that, at least for research in monodisciplinary journals, ChatGPT tends to give higher scores to newer research in all fields, although this tendency is never large and is small in most fields. The results also show substantial disciplinary differences. Together these suggest that the same field normalisation approaches used in citation analysis are also needed for ChatGPT scores: divide the ChatGPT score for any article

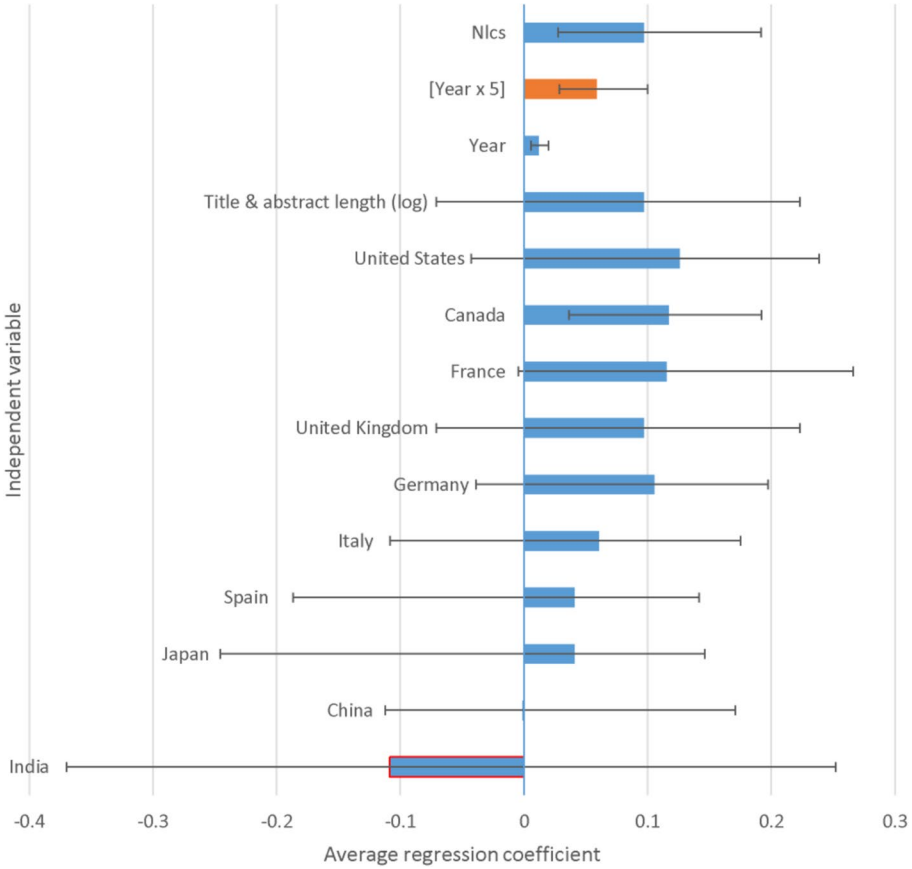

Fig. 9 As Fig. 4 except with article citation rate (NLCS) as an additional independent variable. Dependent variable: ChatGPT score (1–4). Bars represent average regression coefficients across 26 field-specific OLS models; error bars indicate the minimum and maximum values observed across those models. Table 1 (column numbered 9) reports the number of statistically significant coefficients for each dependent variable

by the average ChatGPT score for all articles in the same field and year. This will give a ratio that can be fairly compared between fields and years. For example, a score of 1 would indicate the average ChatGPT score for all articles from the same field and year, whereas a higher score would indicate an above average ChatGPT score.

The first author country differences found could indicate ChatGPT bias and/or underlying international differences in the quality of research, with the latter being widely believed to occur by policy makers. Further research is needed to identify whether both are contributors and, if so, the relative balance between them within each field. Without this, international comparisons based on ChatGPT scores have a degree of uncertainty, as do current citation-based comparisons. In practical terms, it may be better to inform evaluators given ChatGPT scores to support their expert evaluations about this issue and allow them to use their judgement about how far to consider it for the articles that they are reading.

Finally, the abstract length factor found potentially indicates another ChatGPT bias, such as against articles in journals with stricter abstract length restrictions, but, from the discussion, this seems to be primarily due to better (or at least more cited) journals tending

to publish articles with longer abstracts. Nevertheless, more research is needed to investigate the residual abstract length association and decide whether it would ever be appropriate to normalise ChatGPT scores for abstract length in addition to field and year.

Funding This study was supported by an ESRC Metascience Grant UKRI1079.

Declarations

Competing interests The first author is a member of the Distinguished Reviewers Board of this journal.

References

Adams, J. (1998). Benchmarking international research. *Nature, 396*(6712), 615–618.

Bai, X., Wang, A., Sucholutsky, I., & Griffiths, T. L. (2025). Explicitly unbiased large language models still form biased associations. *Proceedings of the National Academy of Sciences, 122*(8), Article e2416228122.

Barrere, R. (2020). Indicators for the assessment of excellence in developing countries. In E. Kraemer-Mbula, R. Tijssen, M. L. Wallace, & R. McClean (Eds.), *Transforming research excellence: New ideas from the Global South* (pp. 219–232). African Minds.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610–623).

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science, 356*(6334), 183–186.

Carbonell Cortés, C., Parra-Rojas, C., Pérez-Lozano, A., Arcara, F., Vargas-Sánchez, S., Fernández-Montenegro, R., & López-Verdeguer, I. (2024). AI-assisted prescreening of biomedical research proposals: Ethical considerations and the pilot case of "la Caixa" Foundation. *Data & Policy, 6*, Article e49.

Chen, S., Arsenault, C., & Larivière, V. (2015). Are top-cited papers more interdisciplinary? *Journal of Informetrics, 9*(4), 1034–1046.

Didegah, F. (2014). *Factors associating with the future citation impact of published articles: A statistical modelling approach*. PhD thesis, University of Wolverhampton.

Dunaiski, M., Geldenhuys, J., & Visser, W. (2019). On the interplay between normalisation, bias, and performance of paper impact metrics. *Journal of Informetrics, 13*(1), 270–290.

Echterhoff, J., Liu, Y., Alessa, A., McAuley, J., & He, Z. (2024). Cognitive bias in decision-making with LLMs. Preprint retrieved from https://arxiv.org/abs/2403.00811

Gov.UK (2022). International comparison of the UK research base, 2022. https://assets.publishing.service.gov.uk/media/628cd2828fa8f55615524e8c/international-comparison-uk-research-base-2022-accompanying-note.pdf

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th annual meeting of the association for computational linguistics*. Association for Computational Linguistics. https://arxiv.org/abs/2004.10964

Hicks, D., Wouters, P., Waltman, L., De Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature, 520*(7548), 429–431.

Holm, J., Waltman, L., Newman-Griffis, D., & Wilsdon, J. (2022). Good practice in the use of machine learning & AI by research funding organisations: Insights from a workshop series. https://orda.shef.ac.

uk/articles/report/Good_practice_in_the_use_of_machine_learning_AI_by_research_funding_organisations_insights_from_a_workshop_series/21710015/1/files/38515406.pdf
Kotek, H., Dockum, R., & Sun, D. (2023). Gender bias and stereotypes in large language models. In *Proceedings of the ACM collective intelligence conference* (pp. 12–24).
Kumar, D., Jain, U., Agarwal, S., & Harshangi, P. (2024). Investigating implicit bias in large language models: A large-scale study of over 50 LLMs. Preprint retrieved from https://arxiv.org/abs/2410.12864
Kumar, C. V., Urlana, A., Kanumolu, G., Garlapati, B. M., & Mishra, P. (2025). No LLM is free from bias: A comprehensive study of bias evaluation in large language models. Preprint retrieved from https://arxiv.org/abs/2503.11985
Langfeldt, L., Nedeva, M., Sörlin, S., & Thomas, D. A. (2020). Co-existing notions of research quality: A framework to study context-specific understandings of good research. *Minerva, 58*(1), 115–137.
Larivière, V., Archambault, É., & Gingras, Y. (2008). Long-term variations in the aging of scientific literature: From exponential growth to steady-state science (1900–2004). *Journal of the American Society for Information Science and Technology, 59*(2), 288–296.
Li, H., Dong, Q., Chen, J., Su, H., Zhou, Y., Ai, Q., & Liu, Y. (2024). LLMs-as-judges: A comprehensive survey on llm-based evaluation methods. Preprint retrieved from https://arxiv.org/abs/2412.05579
Lin, X., & Li, L. (2025). Implicit bias in LLMs: A survey. Preprint retrieved from https://arxiv.org/abs/2503.02776
Martin, B. R. (2016). Editors' JIF-boosting stratagems–Which are appropriate and which not? *Research Policy, 45*(1), 1–7.
Moed, H. F., Markusova, V., & Akoev, M. (2018). Trends in Russian research output indexed in Scopus and Web of Science. *Scientometrics, 116*, 1153–1180.
Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M. E., & Staab, S. (2020). Bias in data-driven artificial intelligence systems: An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10*(3), Article e1356.
Omiye, J. A., Lester, J. C., Spichak, S., Rotemberg, V., & Daneshjou, R. (2023). Large language models propagate race-based medicine. *NPJ Digital Medicine, 6*(1), 195.
Peters, U., & Carman, M. (2024). Cultural bias in explainable AI research: A systematic analysis. *Journal of Artificial Intelligence Research, 79*, 971–1000.
Porter, A., & Rafols, I. (2009). Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. *Scientometrics, 81*(3), 719–745.
Puuska, H. M. (2014). *Scholarly publishing patterns in Finland—A comparison of disciplinary groups*. PhD thesis, University of Tampere.
Rushforth, A., & Hammarfelt, B. (2023). The rise of responsible metrics as a professional reform movement: A collective action frames account. *Quantitative Science Studies, 4*(4), 879–897.
Saad, A., Jenko, N., Ariyaratne, S., Birch, N., Iyengar, K. P., Davies, A. M., Vaishya, R., & Botchu, R. (2024). Exploring the potential of ChatGPT in the peer review process: An observational study. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 18*(2), Article 102946. https://doi.org/10.1016/j.dsx.2024.102946
Technopolis. (2024). REF2021 cost evaluation: Final report. https://repository.jisc.ac.uk/9184/1/REF_2021_cost_evaluation_final_report.pdf
Thelwall, M. (2017). Three practical field normalised alternative indicator formulae for research evaluation. *Journal of Informetrics, 11*(1), 128–151. https://doi.org/10.1016/j.joi.2016.12.002
Thelwall, M. (2024). Can ChatGPT evaluate research quality? *Journal of Data and Information Science, 9*(2), 1–21. https://doi.org/10.2478/jdis-2024-0013
Thelwall, M. (2025). Evaluating research quality with large language models: An analysis of ChatGPT's effectiveness with different settings and inputs. *Journal of Data and Information Science, 10*(1), 7–25. https://doi.org/10.2478/jdis-2025-0011
Thelwall, M., Kousha, K., Stuart, E., Makita, M., Abdoli, M., Wilson, P., & Levitt, J. (2023). In which fields are citations indicators of research quality? *Journal of the Association for Information Science and Technology, 74*(8), 941–953. https://doi.org/10.1002/asi.24767
Thelwall, M., & Yaghi, A. (2024). In which fields can ChatGPT detect journal article quality? An evaluation of REF2021 results. Preprint retrieved from https://arxiv.org/abs/2409.16695
Traag, V. A., & Waltman, L. (2022). Causal foundations of bias, disparity and fairness. Preprint retrieved from https://arxiv.org/abs/2207.13665
Waltman, L., & Van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology, 63*(12), 2378–2392.

Wang, J. (2013). Citation time window choice for research impact evaluation. *Scientometrics, 94*(3), 851–872.
Wang, Z., Wu, Z., Zhang, J., Jain, N., Guan, X., & Koshiyama, A. (2024). Bias amplification: Language models as increasingly biased media. Preprint retrieved from https://arxiv.org/abs/2410.15234
Wilhite, A., Fong, E. A., & Wilhite, S. (2019). The influence of editorial decisions and the academic network on self-citations and journal impact factors. *Research Policy, 48*(6), 1513–1522.
Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., Hill, S., & Johnson, B. (2015). The metric tide: Independent review of the role of metrics in research assessment and management. https://www.ukri.org/publications/review-of-metrics-in-research-assessment-and-management/
Wilson, K., & Caliskan, A. (2024). Gender, race, and intersectional bias in resume screening via language model retrieval. In *Proceedings of the AAAI/ACM conference on AI, ethics, and society* (Vol. 7, pp. 1578–1590).
Ye, R., Pang, X., Chai, J., Chen, J., Yin, Z., Xiang, Z., & Chen, S. (2024). Are we there yet? Revealing the risks of utilizing large language models in scholarly peer review. Preprint retrieved from https://arxiv.org/abs/2412.01708
Zhao, Y., Wang, B., & Wang, Y. (2025). Explicit vs. implicit: Investigating social bias in large language models through self-reflection. Preprint retrieved from https://arxiv.org/abs/2501.02295