
Article:
Thelwall, M. orcid.org/0000-0001-6065-205X and Nunkoo, R. (2025) Evaluating the quality of tourism research using ChatGPT. Current Issues in Tourism. ISSN: 1368-3500

https://doi.org/10.1080/13683500.2025.2596265

EVALUATING THE QUALITY OF TOURISM RESEARCH USING CHATGPT

Mike Thelwall
Information School,
University of Sheffield, UK.
https://orcid.org/0000-0001-6065-205X
m.a.thelwall@sheffield.ac.uk

*Robin Nunkoo, PhD [1,2,3,4]
[1] Department of Management, University of Mauritius, Reduit MU 80837, Mauritius
[2] School of Tourism and Hospitality, University of Johannesburg, South Africa
[3] Kyung Hee University, Seoul, Republic of Korea
[4] School of Hospitality, Tourism & Events; Centre for Research and Innovation in Tourism (CRiT), Taylor's University, Malaysia.
Email: r.nunkoo@uom.ac.mu

Abstract
This article evaluates the quality of tourism research by comparing the results obtained from citation analysis with those obtained from ChatGPT quality scores based on the Research Excellence Framework criteria, applied to 50 of the largest tourism and leisure journals. Whilst there was an article-level weak correlation between normalized citation rates and normalized ChatGPT scores, there was a moderately strong journal-level correlation between the two measures. The latter supports the value of normalized ChatGPT scores for journal quality indicators. The results also suggest important nuances between and within qualitative and qualitative methods used in tourism. Articles using experiments, advanced statistical tests, and theories scored well on both indicators but the opposite for those based on surveys and convenience sampling.
Keywords: ChatGPT, Large Language Models, bibliometrics, citation analysis, research evaluation.

1. Introduction

The volume of academic research in tourism has risen steadily since the 1980s (Benckendorff & Zehrer, 2013). Tourism is now an established field of intellectual inquiry, characterized by a strong scientific community of researchers and multiple publication outlets. Tourism scholars engage in research for reasons including advancing knowledge, career progression, meeting job requirements, and personal satisfaction, and to demonstrate academic leadership in the field (Law & Chon, 2007). They therefore need to assess the quality of research in their field to support decisions about which topics are worth researching, who to collaborate with, and even which journals to submit to. Research managers also need to know this to help with appointments, promotion, and tenure decisions. Unfortunately, evaluating published academic research quality is a difficult, time consuming, and partly subjective task. For this reason, tourism scholars often use citation-based indicators as proxies for expert evaluations or to support them. For example, researchers might use Journal Impact Factors to help decide which journal to publish in and employers might consult the citation record of applicants when shortlisting for a new post. Another use for citation-based indicators is to support analyses of departmental research for formative self-evaluations (Nunkoo, Hall, Rughoobur-Seetah, & Teeroovengadum, 2019).

In addition to the scholarly impact of scientific articles, research quality is also sometimes assessed by the extent to which tourism research improves the lives of people, with a focus on the socio-economic benefits, including public engagement, as part of national assessments such as the UK's Research Excellence Framework (Phillips, Page, & Sebu, 2020). Unfortunately, citation-based indicators have several limitations in this respect. In addition to being opened to manipulation (Ioannidis, 2024), citations primarily reflect scholarly impact, whereas societal impact, originality and rigor are also widely considered to be core components of research quality (Langfeldt et al., 2020). Whilst carefully constructed citation-based indicators correlate positively with research quality in all fields, the strength varies from very weak (most arts and humanities, some social sciences) to moderately strong (e.g., health, life and physical sciences) (Thelwall et al., 2023). In practice, they are too weak in many fields to be useful, and they seem to generate an overreliance on citation data (e.g., sfdora.org). Another important problem is that they are useless for research that is less than three years old because citation data needs this long to mature (Wang, 2013), and recent research can be the most important for decision making. It is for these reasons that some tourism researchers question the value of citation-based metrics, arguing for a change in the evaluation of tourism research (Benjamin, Lee & Boluk, 2024; Dolnicar, 2025).

ChatGPT has recently emerged as an alternative to citations as a source of evidence about the quality of journal articles (e.g., Li & Qiu, 2024). If it is fed with instructions about how to evaluate research quality and asked to score a journal article from its title and abstract then the results correlate positively with independent research quality measures or indicators in all broad fields and correlate more strongly than citation-based indicators in most (Thelwall et al., 2024; Thelwall & Yaghi, 2025). There are many other large language models, but none have shown a superior ability to ChatGPT for this task yet. Although the field of tourism and leisure has not been independently tested, the ChatGPT predictions correlate much more strongly with indicators of research quality than citation-based indicators for the combined Research Excellence Framework (REF) field of Sport and Exercise Sciences, Leisure and Tourism (Thelwall & Yaghi, 2025). This suggests that ChatGPT scores might give a much stronger research quality indicator for tourism and leisure than citations. In this context, the current study addresses three research gaps. First, it assesses for the first time the value of ChatGPT for tourism and leisure, in case the relationship for the encompassing REF field is

weak for this subset, or in case the relationship is UK-specific rather than international. Second, it identifies for the first time the types of articles in the field that ChatGPT gives high or low scores to, so that the implications of relying on ChatGPT rather than citation data could be assessed. Third, the availability of quality scores for a large set of Leisure and Tourism articles gives an opportunity to reflect on the types of research that the field might consider to be particularly valuable, rather than just more citable.

In the context of the need to assess the use of ChatGPT for evaluating the tourism and leisure field, the current study analyses recent articles published between 2011 and 2020 (n = 19,426) in the 50 largest tourism, hospitality, and leisure journals, comparing the results to those obtained through citation analysis for the same dataset. The comparison with citations for older research, although ChatGPT may be superior, especially for newer studies, is needed to help validate ChatGPT scores as a research quality indicator, given the absence of a large set of expert quality score judgements for tourism and leisure research that would allow a direct test. Following the UK's Research Excellence and the most used quality criteria in the Global North (Langfeldt et al., 2000), this study defines research quality as comprising originality, significance, and rigor. The following research questions drive the study, motivated by the ways that citations might be used to support expert judgement in the field: RQ1: Does the citation rate of a tourism and leisure article correlate with its ChatGPT research quality score?; RQ2: Does the average citation rate of a tourism and leisure journal correlate with its average ChatGPT research quality score?; and RQ3: Are there any patterns in the article types or topics that tend to get higher ChatGPT research quality scores, and, if so, are they different from the patterns in the types of articles that tend to be most cited?

By answering these questions, the study goes beyond an assessment of the citation impact of tourism research by considering its broader impact on society and economy, as well as its originality and rigor, as defined by the UK Research Excellence Framework, using a novel Large Language Model (LLM) tool, ChatGPT. In so doing, we address the limitations of data based on productivity and citations for indicators of research quality and contribute to the ongoing debates on social impact assessments of tourism research (e.g., Brauer, Dymitrow, & Tribe, 2019; Viana-Lora, 2023). We also address the need to develop alternative research assessment parameters that consider a broader set of criteria such as journals' relevance and significance (Jamal, Smith, & Watson, 2008).

2. Research Assessments in Tourism

The maturation of tourism research means that it is under constant evaluation, with an emphasis on research quality. However, what constitutes research quality is highly debated and contested in tourism (Brauer et al., 2019; Phillips et al., 2020). Bibliometric analysis involving the ranking of institutions, journals, and authors has been the most common approach to evaluating tourism research (Hall, 2011). One approach has been to identify the most prolific tourism authors based on their numbers of publications (e.g., Ryan, 2005). These rankings confer power and prestige to these entities and, therefore, significant competitive advantage (Buckley, 2019). Of course, research publishing volume also does not equate to impact (Phillips et al., 2020). It also shifts scholars' focus from engaging in theoretically and practically relevant research to publishing large numbers of articles, irrespective of their value to literature and society.

Citation analysis is another frequently used method to evaluate tourism research. McKercher (2005), for example, used Google Scholar citation data to identify the most cited tourism

scholars for the period 1970 to 2007. Other studies have investigated the determinants of tourism articles' citations (e.g., Polat, Celik, Arici, & Koseoglu, 2024). Citation data also forms the basis of tourism journal rankings such as the JIF (Clarivate), Cite Score (Elsevier), and SCImago Journal Rank (SJR). Evaluating research using citation impact is not without caveats. In addition to being influenced by several author-related characteristics, such as gender, geographic location, institutional affiliation, and reputation, citation indices favor certain types of tourism research and methodologies over others (Nunkoo et al., 2019). In their assessment of 27 broad fields as classified by Scopus, another study found that time was also a factor because the citation impact of research based on interviews and focus groups decreased over time (Thelwall & Nevill, 2021). Citations can also be manipulated to artificially increase scholarly impact when, for example, researchers engage in excessive self-citations and form citation cartels to promote their works by citing each other inappropriately (Bartneck & Kokkelmans, 2011). Journals have also been accused of manipulating citations to increase impact factors (Ioannidis & Maniadis, 2024).

2.1. Research Quality: Originality, Rigor, and Significance

In discussions on the unintended consequences of metric-based evaluations on the production of tourism knowledge, it is clear that research quality cannot be assessed primarily using number of articles and citation-based indices (e.g., Benjamin et al., 2024; Dolnicar, 2025). Benjamin and Lee (2023) even assert that the metrication of tourism knowledge based on research outputs and citations will lead to "the death of tourism scholarship". In response to this critique and increased accountability in the use of state funding for research, requiring the scientific community to demonstrates the 'non-academic' impacts of their research, a new paradigm of research quality based on the relevance of tourism research for societies and economies has emerged (Thomas, 2024; Viana-Lora, 2023).

National research assessments exercises such as the UK Research Excellence Framework manifest the need to assess the value of scientific research in improving the lives of people, societies, and economies (Brauer et al., 2019). For the first time in 2014, in addition to evaluating research based on its originality and rigor, the UK Research Excellence Framework introduced a third criterion of research quality - significance (Brauer et al., 2019; Phillips et al., 2020). Significance is "the extent to extent to which the work has influenced, or has the capacity to influence, knowledge and scholarly thought, or the development and understanding of policy and/or practice". Originality is defined as "the extent to which the research makes a significant and original contribution to understanding and knowledge in the field". Rigor is "the extent to which the work demonstrates intellectual coherence and integrity, and adopts robust and appropriate concepts, analyses, sources, theories and/or methodologies" (REF, 2021a, p. 34-35).

The criteria of originality, rigor, and significance of tourism research have been discussed in literature, although not always in the context of research assessment exercises. Sánchez, Makkonen, and Williams (2019), for example, assessed originality by interviewing editors and editorial board members of tourism journals. Based on the findings, they developed a typology of originality based on the contributions of the research to tourism knowledge: minor incremental, incremental (testing existing theories in new contexts), major incremental (new to tourism studies/only original in the field), and radical (purely original regardless of the field). Studies have also focused on the drivers and inhibitors of originality in tourism research, such as the individual traits of the researcher, discipline and network, and the research system (e.g., Sanchez, Mantecón, Williams, Makkonen, & Kim, 2022) and provided

guidelines to develop research that makes an original contribution to tourism knowledge (e.g., Tribe, 2018). However, there is a consensus that originality in tourism research is hard to define and assess (Buckley, 2023).

Rigor in tourism research has been implicitly addressed in discussions of theory and methodology in tourism research. Several scholars discuss methodological rigor in qualitative and quantitative tourism research (e.g., Riley & Love, 2000) or in specific methods tourism researchers use such as sampling (e.g., Czernek-Marszałek & McCabe, 2024), experimental design (e.g., Leung, Fong, Xue, & Mattila, 2024), formative and reflective measures (e.g., Olaru & Hofacker, 2009), and statistical techniques (e.g., Ali, Rasoolimanesh, Sarstedt, Ringle, & Ryu, 2018). At the theoretical level, tourism research is criticized for lacking rigor because researchers engage passively with theory, use theory loosely and misleadingly, and wrongly consider diagrams, statistical analysis, and concepts as theory (McCabe, 2024; Nunkoo & Armbrecth, 2025). Rigor in tourism has also been discussed in relation to writing and ethics (Tribe, 2018).

The significance of tourism research has been addressed in some recent studies. For the Research Excellence Framework, non-academic impact (a component of significance) is impact manifesting outside academia and does not include scholarly impact measured by citations and impact through teaching (Brauer et al., 2019). It is in this context that studies have used information submitted by universities to the UK Research Excellence Framework to examine the pathways through which tourism research impacts society and economy, identifying the 'significance gap' and the resulting implications for universities and tourism scholars (Phillips et al., 2020; Tribe & Paddison, 2024). Thomas and Ormerod's (2017) empirical assessment of the impact of tourism research on policy and practice suggests that tourism research has little influence on practice compared to other social sciences. Viana-Lora, Nel-lo-Andreu, and Anton-Clavé (2023) adopt a more conceptual approach by developing a framework to assess the social impacts of tourism research. They advocate for methods that measure the impact of tourism research on societies and economies.

3. Data and Methods

The research design was to collect a large recent collection of tourism and leisure journal articles and then correlate the ChatGPT scores with the citation rates by article and by journal, investigating anomalies and using a word frequency method to investigate articles with high and low ChatGPT scores and/or citation rates (Figure 1, Table 1). The initial dataset consisted of all records for the Tourism, Leisure and Hospitality Management category in Scopus and published between 2011 and 2020. These were downloaded from Scopus between January and March of 2024. The range 2011-2020 contains ten years, supporting a fine-grained analysis of topics (see below), and all articles have had at least three full years to attract citations, which is usually sufficient for citation analysis (Wang, 2013) for the comparisons. Scopus was chosen rather than the Web of Science for slightly wider coverage and instead of Dimensions for clearer journal-based categories.

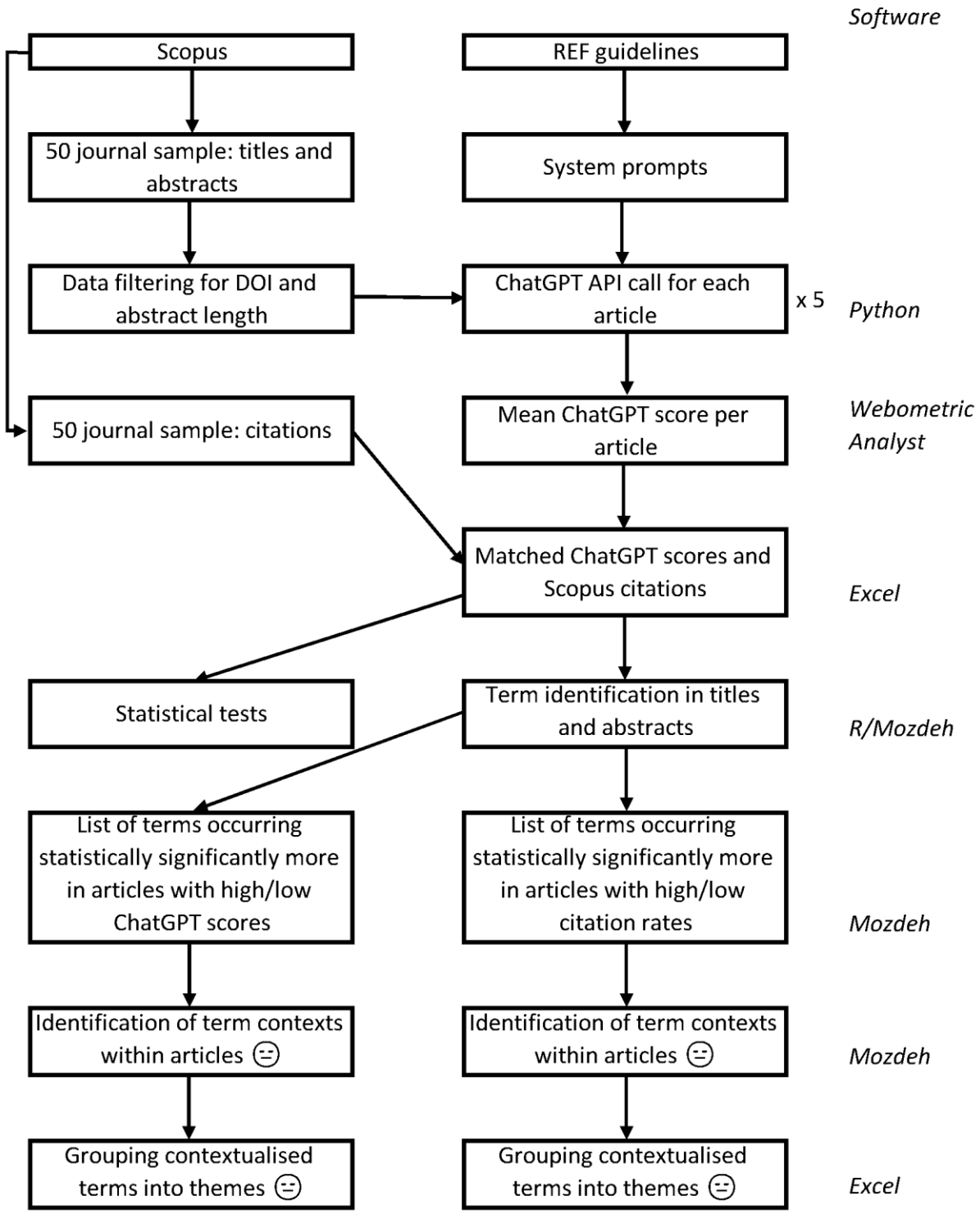

Figure 1. A flow diagram of the research process. Faces indicate subjective analyses and the main software used is shown on the right.

Table 1. The main variables used in the study.

Variable	Meaning
ChatGPTn	Average of five scores given to an article by ChatGPT 4o-mini, corrected for the publication year by subtraction so the mean ChatGPTn.is the same for all years.
ChatGPTz	Average of five scores given to an article by ChatGPT 4o-mini, corrected for the publication year by z-normalization so that the ChatGPTz scores for each year have a mean of 0 and a standard deviation of 1.
NLCS	Normalised Log Citation Score – field and year normalized citation rate.
MNLCS	The mean NLCS for all articles in the journal from this study.
[Year]	The official publication year of the journal article. Used to calculate ChatGPTn
[Journal]	The official journal publishing the article. Used for the leave-one-out analysis.

The dataset was cleaned to make it more homogeneous, as follows. First, articles without DOIs were removed since DOIs were needed for duplicate checking and some of these were not full articles. Second, articles with short abstracts were removed, using 730 characters as a cut-off, which equates to removing 10% of the articles. These might be editorial material, short articles, or indicative of an error, such as an accidentally truncated abstract. Removing these is necessary to make comparisons between ChatGPT and citation indicators fairer. Previous studies with ChatGPT have also removed articles with short abstract, each using a heuristically determined cutoff or 10% (e.g., Thelwall, 2025b; Thelwall & Yaghi, 2025). The value of 10% was chosen as a default. It makes the evaluation fairer by removing articles where the abstract is less likely to give an adequate summary. The article with the largest abstract was also removed after manual checks showed that it was incorrect. Next, articles with duplicate DOIs were removed if the articles were different (because an error must have occurred), or the incorrect article was removed when one was an early version and the other was the final version. Finally, when a page range was reported in Scopus, articles with less than six pages were removed since these may be short submissions, such as brief communications or editorial material. The six-page limit was heuristically determined by sorting the list of articles by page length (when recorded) and noticing that there were relatively few articles with less than 6 pages and these seemed to be small contributions, even though not officially classified as brief communications or similar.

Some of the 143 journals in Scopus's Tourism, Leisure and Hospitality Management category were of relevance to tourism and leisure but mainly covered other topics so the list was filtered to just the journals mentioning tourism, hospitality, leisure or equivalent terms (see Appendix 1). This excluded journals primarily about, for example, cities, heritage, sport and business, but not when they also combined elements of leisure or tourism. Only the largest 50 journals (in terms of the number of articles in the period) were kept, so that the comparison between journals would not be complicated by small journals with unreliable summary statistics (i.e., wide confidence intervals). The smallest of the 50 journals had 138 articles in the filtered dataset. Of course, the relationship between ChatGPT scores and citations may be different for smaller journals and non-Scopus journals, but the focus of the current paper is on the international core of leisure and tourism research, as represented by these journals.

3.1. ChatGPT scores

The combined title and abstract of each article in the dataset was sent to the ChatGPT 4o-mini API five times to obtain a research quality score. With each request, system instructions

were sent that are the UK Research Excellence Framework guidelines for social science expert reviewers (Main Panel C), slightly rephrased to adapt them for the ChatGPT instruction style and used previously by Thelwall and Yaghi (2025) (Appendix 2). Despite being designed for the UK, the Research Excellence Framework guidelines are broadly appropriate because they are based around the widely agreed core components of impact, originality and rigor (Langfeldt et al., 2020), include detailed criteria for research quality judgements, and define four different levels: 1* "nationally recognized", 2* "internationally recognized", 3* "internationally excellent" and 4* "world leading", so give an evidenced score on a defined numerical scale. The articles were sent five times because it is common for an article to receive different scores from ChatGPT due to the inbuilt randomness in its algorithm and previous studies have shown that averaging multiple submissions gives improved results (Thelwall, 2024, 2025; Thelwall & Yaghi, 2025). System prompts and ChatGPT parameters (e.g., temperature) were not varied because previous research had found alternatives to give the same or worse results; similarly, entering the title alone or the full text gave worse results (Thelwall, 2025a). Note that the system was only fed with the title and abstract and not any other information (e.g., no authors, no publication venue, no publication year, no DOI). The queries were submitted on the 6th of November 2024 to the latest ChatGPT 4o-mini at the time (gpt-4o-mini-2024-07-18) in batch mode. Source code for the API is available at https://github.com/MikeThelwall/LargeLanguageModels and the program for extracting the scores from the reports is at https://github.com/MikeThelwall/Webometric_Analyst.

ChatGPT scores tended to be slightly lower for earlier years, presumably because older articles seemed less novel, so they were normalized by adding or subtracting an amount that would make the average ChatGPT score for each year equal to the overall average, 2.633. This normalized value is denoted ChatGPTn. We chose this simple approach rather than a more sophisticated version, such as z-normalization, to give a transparent and understandable approach. As a robustness check we also ran the tests with the raw ChatGPT scores and with z-normalized scores (subtracting year ChatGPT mean from the ChatGPT score and then dividing by the year ChatGPT standard deviation). Both year normalisation approaches tend to reduce the most common scores, as shown by their histograms (Freedman & Diaconis, 1981) and kernel density estimates (Sheather & Jones, 1991) (Figures 2, 3, 4).

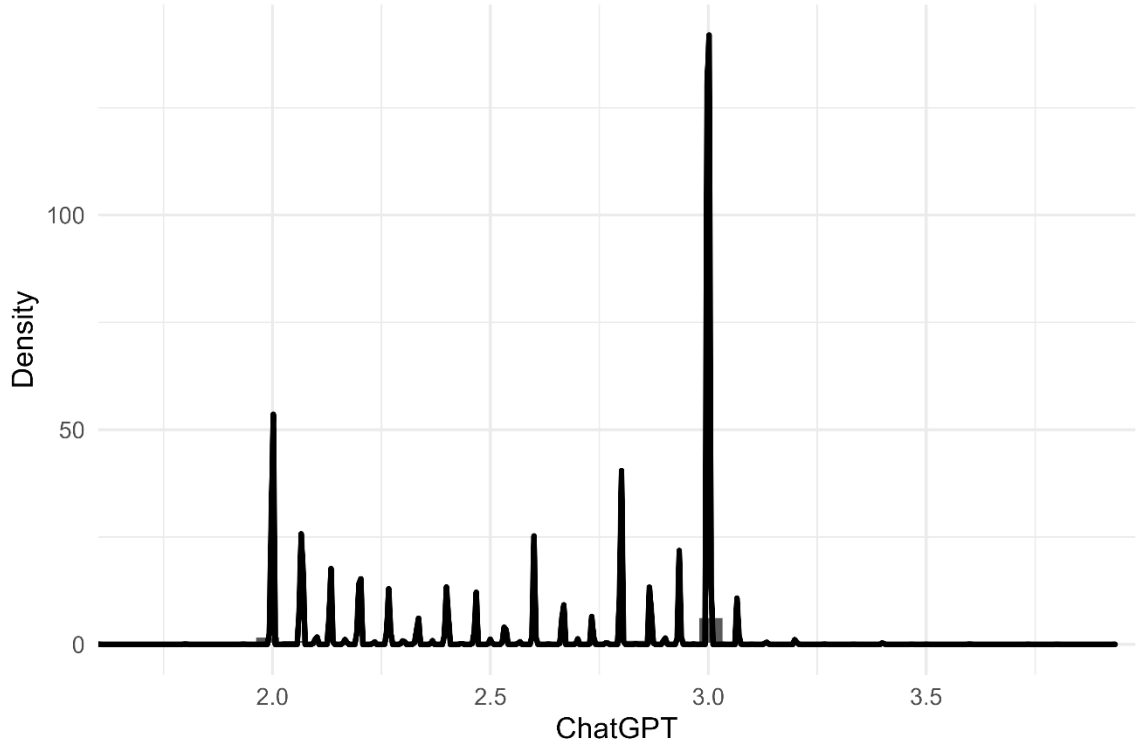

Figure 2. A histogram and kernel density estimate of the distribution of the raw ChatGPT scores (average of 5 per article).

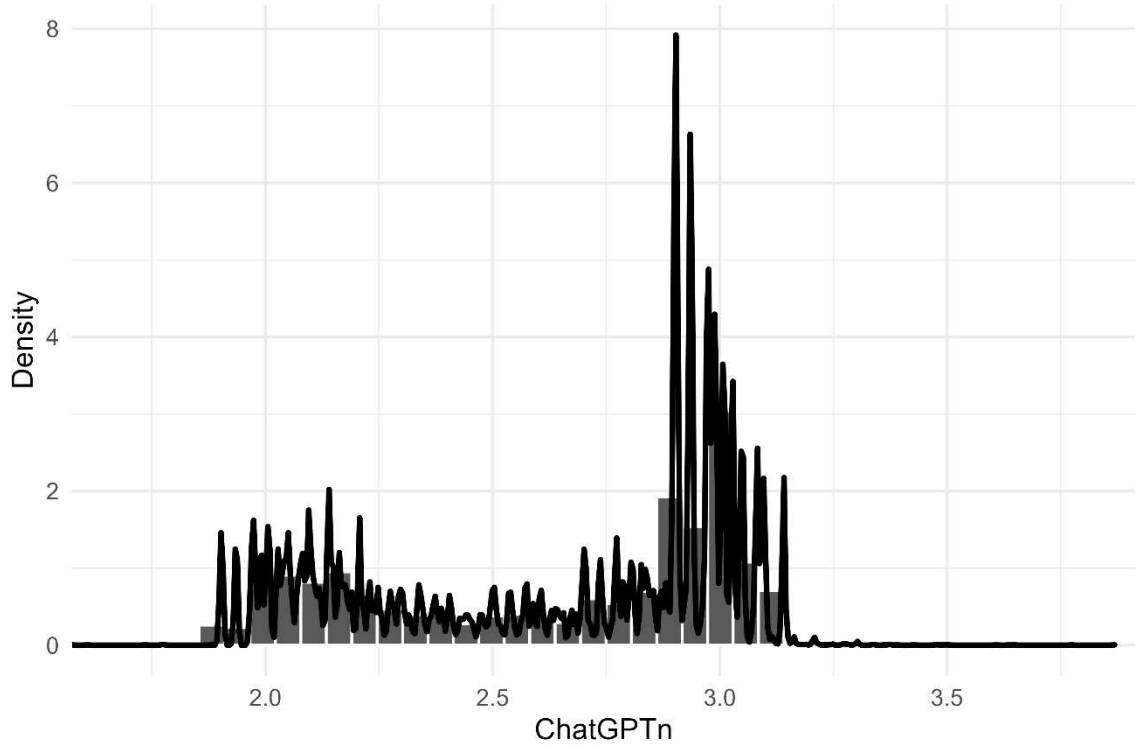

Figure 3. A histogram and kernel density estimate of the distribution of year normalized ChatGPT scores.

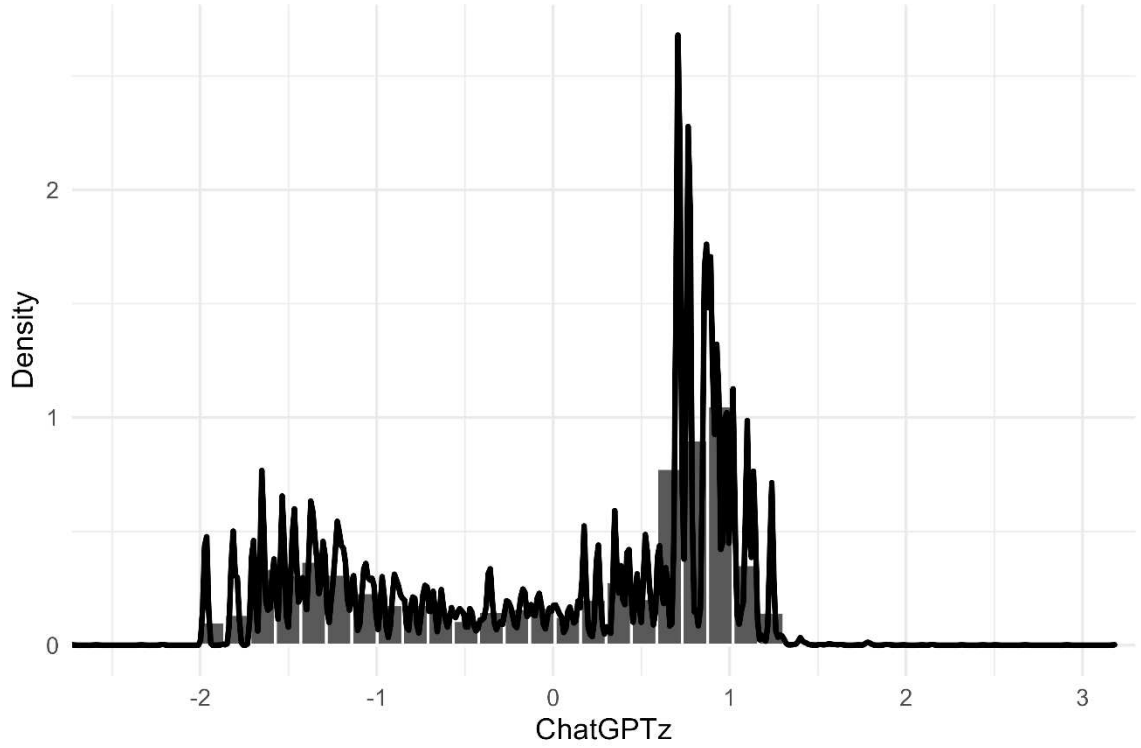

Figure 4. A histogram and kernel density estimate of the distribution of z-normalized ChatGPT scores.

3.2. Citation data

A field and year normalized citation count, the Normalized Log-transformed Citation Score (NLCS) was used for each article as an unbiased (for publication year) citation impact estimate. For this, the raw citation count, for each article was replaced by log(1+citation_count) to reduce the skewness in the citation dataset. The log(1+citation_count) value for each article was then divided by the mean log(1+citation_count) for all articles from the same year in the dataset, giving the NLCS value. This is, by construction, fair to compare between years because the mean NLCS is 1 in all years. The average citation impact of each journal was calculated as the mean of the NLCS values of all articles in that journal. This is known as the Mean NLCS or MNLCS (Thelwall, 2017). This is a reasonable indicator of average citation impact because it is not unduly influenced by individual highly cited articles (because of the log transformation) and does not advantage older articles that have had more time to be cited. Note that self-citations were not excluded from the data.

3.3. Analysis

For RQ1, the ChatGPT scores were correlated with citation rates for the articles to see how closely the two associate. For RQ2, the mean ChatGPTn scores were correlated with average citation rates (MNLCS) for journals to see how closely the two associate. Spearman correlations were used for the primary statistical analysis because the rank order of the articles or journals is the key variable of interest. Within-journal and leave-one-journal-out analyses were included for robustness, and bootstrap 95% confidence intervals were calculated. Pearson correlations were used as an additional robustness check and are valid

because the transformations had reduced the skewing in the dataset and both variables passed a Shapiro-Wilk test for being plausibly normal (i.e., not rejecting the null hypothesis that the data is normally distributed). To distinguish between journal-level and article-level effects, within journal correlations were calculated and multilevel modeling applied, with NLCS as the dependent variable and the formula NLCS ~ ChatGPTn + (1 + ChatGPTn | Journal).

For RQ3, Word Association Thematic Analysis (WATA: Thelwall, 2021, 2023) was applied to identify terms that tended to occur disproportionately often in articles with high ChatGPT scores. WATA works by first identifying terms that occur statistically significantly disproportionately often in one dataset (e.g., high scoring articles) compared to another (e.g., low scoring articles). The threshold p=0.001 was used for this to reduce the chance of spurious positives. The free software Mozdeh (https://github.com/MikeThelwall/Mozdeh) was used for this.

The next stage, term contextualization, involves reading a sample of the texts (titles/abstracts) containing each identified term and recording its typical context. For example, the context of "mediate" might have been "intercede between angry tourists", or "resolve a business dispute" but was "a statistical variable relating to other variables". Finally, the term contexts were iteratively and reflexively clustered into themes of related terms with a thematic analysis approach. For example, "mediate" with the above context was grouped with other terms including (structural equation) "modelling" into a *complex statistics* theme. The WATA method was repeated for articles with low ChatGPT scores, articles with high citation rates and articles with low citation rates, comparing the results qualitatively.

WATA was used instead of traditional content analysis or thematic analysis because it focuses on finding differences between two collections of text. It was used in preference to variants of topics analysis because WATA is backed by a statistical tests, only including terms that have statistically significant differences ($p<0.001$) between the two sets after a Benjamini-Hochberg familywise error rate correction. The context detection and thematic analysis at its core also supports the face validity of the results. Its weaknesses are the subjective nature of the interpretations and the potential for false positives in the word frequency tests due to different topics frequently employing the same terms.

4. Results

RQ1: Does the citation rate of a tourism and leisure article correlate with its ChatGPT research quality score?

For the 19,426 articles from 2011-2020 in the 50 journals, the average normalized ChatGPT score ChatGPTn correlated moderately and positively with the article normalized citation rate NLCS (Spearman's rho=0.2604, Confidence Interval: [0.2459,0.2737]), indicating a weak positive relationship between the normalized ChatGPT score and citation rates. The leave-one-journal-out analysis also gave quite similar results, so the correlation is not due to a single journal. The results were also similar for Pearson correlations, so both types of correlation suggest a similar strength association (Table 2, rows 2-5).
All results for ChatGPTn were similar to the corresponding results for the z-normalized ChatGPTz (Table 2, rows 9-12). Thus, the two approaches to score normalization seem to be statistically equivalent.

Table 2. Correlation tests on the articles (n=19426 except for the leave-one-journal-out tests).

Score	Dataset	Spearman	L95	U95	Pearson	L95	U95
ChatGPTn	Leave one out min	0.2108	0.1974	0.2249	0.2275	0.2139	0.2411
ChatGPTn	Leave one out max	0.2679	0.2540	0.2812	0.2913	0.2782	0.3042
ChatGPTn	Leave one out mean	0.2603			0.2844		
ChatGPTn	All articles	0.2604	0.2459	0.2737	0.2846	0.2716	0.2975
ChatGPTn	Within journal min	-0.0086			-0.0087		
ChatGPTn	Within journal max	0.3394			0.3119		
ChatGPTn	Within journal mean	0.1168			0.1233		
ChatGPTz	Leave one out min	0.2116	0.1983	0.2256	0.2282	0.2146	0.2417
ChatGPTz	Leave one out max	0.2688	0.2550	0.2819	0.2929	0.2798	0.3059
ChatGPTz	Leave one out mean	0.2612			0.2860		
ChatGPTz	All articles	0.2613	0.2469	0.2744	0.2862	0.2733	0.2991
ChatGPTz	Within journal min	-0.0069			-0.0105		
ChatGPTz	Within journal max	0.3369			0.3090		
ChatGPTz	Within journal mean	0.1167			0.1234		
ChatGPT	Leave one out min	0.2257	0.2121	0.2395	0.2314	0.2178	0.2449
ChatGPT	Leave one out max	0.2777	0.2640	0.2914	0.2890	0.2759	0.3020
ChatGPT	Leave one out mean	0.2706			0.2823		
ChatGPT	All articles	0.2707	0.2573	0.2838	0.2825	0.2695	0.2954
ChatGPT	Within journal min	-0.0241			-0.0237		
ChatGPT	Within journal max	0.3007			0.2937		
ChatGPT	Within journal mean	0.1166			0.1241		

The Spearman correlations for ChatGPT (without normalization, Table 2, rows 16-19) tended to be higher than the corresponding results for ChatGPTz and ChatGPTn, although the difference is not statistically significant (e.g., the means are inside each other's 95% confidence intervals). Thus, the year normalization does not seem to be necessary and may even be counterproductive. This may be a side-effect of a high number of identical scores (e.g., 3*) that any field normalization would shift marginally for each year. This conjecture is supported by the Pearson correlations (which are not rank-based) being higher for the year normalized ChatGPTz and ChatGPTn.

The correlations between citation rates and ChatGPTn scores within individual journals are generally lower, ranging from rho=-0.0086 to 0.3394 with a mean of 0.1168 and from r=-0.0087 to 0.3119, with a mean of 0.1233 (Table 2, rows 6-8). Thus, the overall correlations are primarily due to differences between journals rather than to differences within journals. This is expected because journals in tourism are assumed to have different average quality levels as well as different citation rates. Nevertheless, it also shows that differences between journals alone are insufficient to account for the correlation. The importance of journals is supported by the multilevel modelling analysis (Table 3), where the Intraclass Correlation Coefficient (ICC) of 0.4334 suggests that 43% the variance lies between the journals rather than within them.

Table 3. Multilevel modelling tests on the articles (n=19426).

Variable	Model	Sigma residual	ICC
ChatGPTn	NLCS ~ ChatGPTn + (1 + ChatGPTn | Journal)	0.2893	0.4334
ChatGPTz	NLCS ~ ChatGPTz + (1 + ChatGPTz | Journal)	0.2893	0.3003
ChatGPT	NLCS ~ ChatGPT + (1 + ChatGPTn | Journal)	0.2893	0.4156

The overall ChatGPT average for articles is relatively low at 2.628 on the four-point scale and very few articles score above 3.2 or below 1.8 so a relatively narrow range is used (Figure 5). A Grubb's test suggests that there are no outliers in the data from the perspective of the normal distribution. Qualitatively, however, uncited articles (NLCS=0) with ChatGPTn >3 are anomalies, as are highly cited papers (NLCS>1.5) with ChatGPTn <2. For these, either ChatGPT or citations may be poor indicators of their value. Articles scoring above 3.5 or below 1.8 for ChatGPTn are also unusual. A manual examination of the lowest scoring articles found that in some cases they were clearly weak from a research quality perspective. For example, the minimum score was given to an article that was a personal reflection of the author's lifetime as a researcher in the field. In contrast the highest scoring (GPTn=3.87) was Benali and Ren's (2019) article combining theory and empirical data in an original way: "Lice work: Non-human trajectories in volunteer tourism".

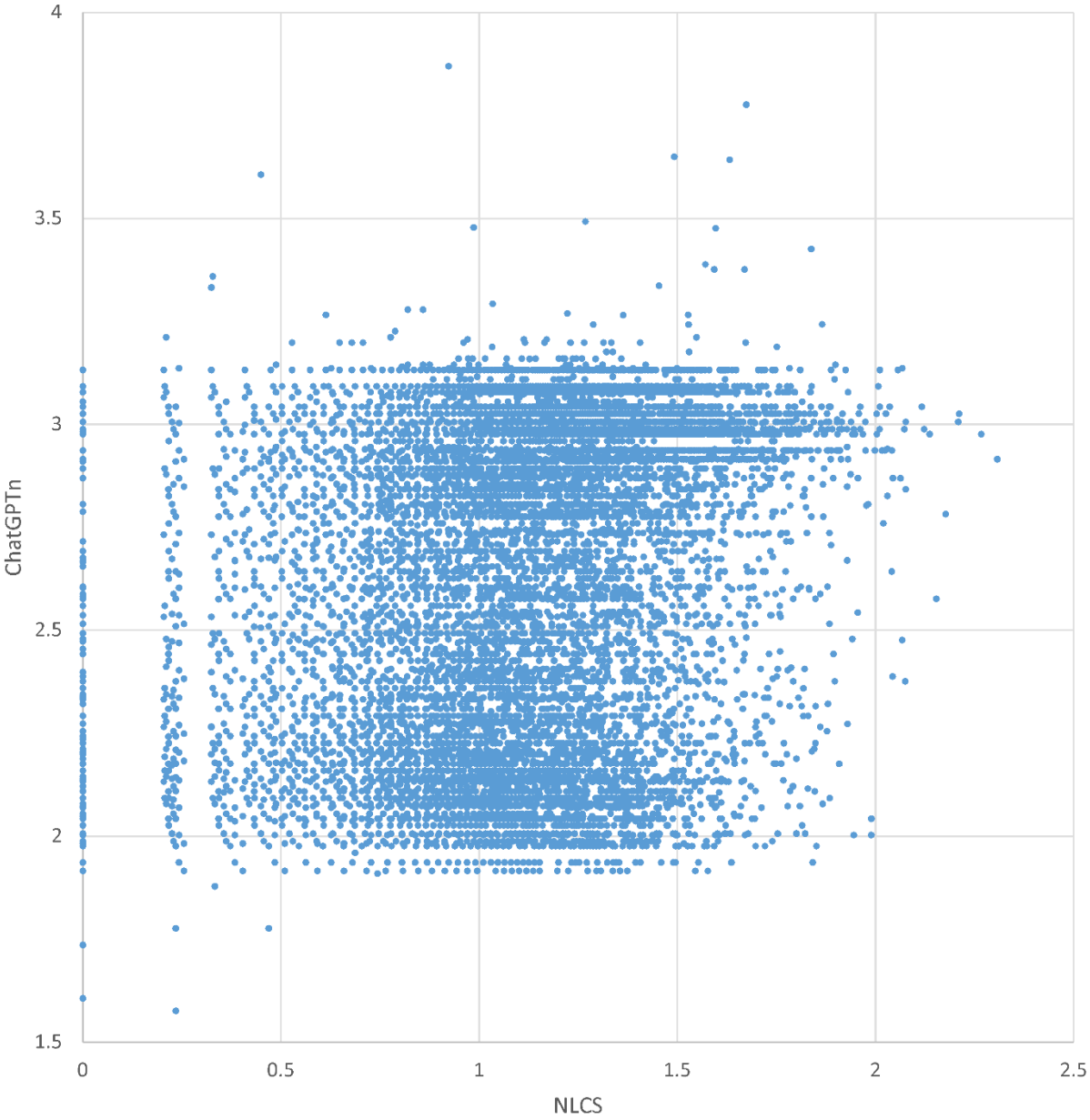

Figure 5. Normalized ChatGPT score against NLCS year-normalized citation rates for non-short articles with DOIs published 2011-2020 (n= 19,426).

RQ2: Does the average citation rate of a tourism and leisure journal correlate with its average ChatGPT research quality score?

For the largest 50 journals, the journal average normalized ChatGPT score correlated strongly, statistically significantly, and positively with the journal normalized citation rate MNLCS (r=0.598, 95% Confidence Interval: [0.384, 0.752], n=50) (Figure 6). Journals towards the top left-hand side of the graph tend to have relatively high ChatGPT scores for their citation rates: leisure journals and those with "Studies" in their names seem overrepresented here.

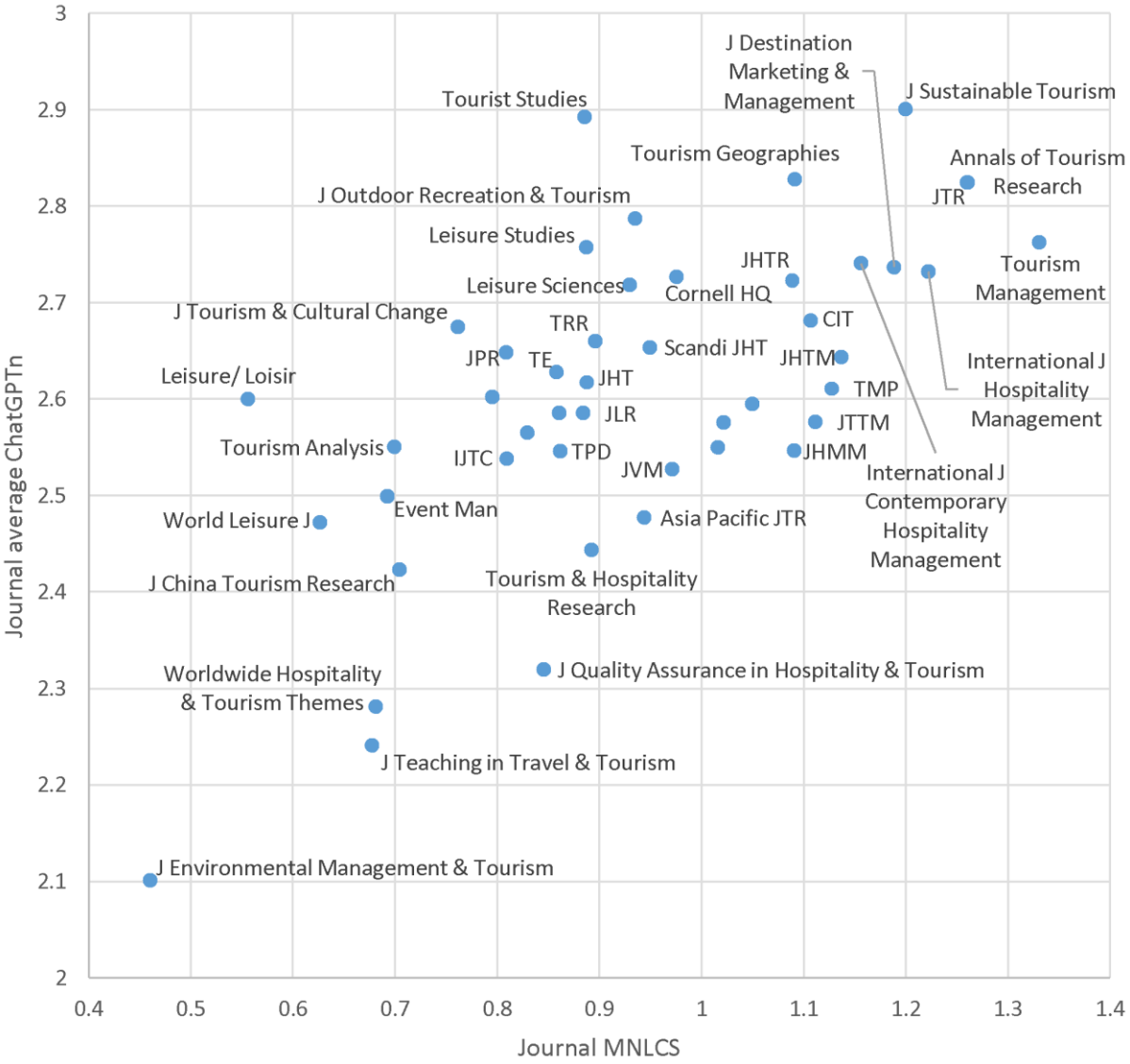

Figure 6. Journal average normalized ChatGPT score against journal MNLCS average citation rates for non-short articles with DOIs published 2011-2020. See Appendix 1 for exact values and sample sizes.

RQ3a: Themes for higher ChatGPT research quality scores and/or more cited articles

Many themes were identified by WATA in articles with higher ChatGPT scores and/or more citations, and these have been clustered into groups to help understand them. The general themes (Table 4) are discussed separately from the topic-based themes (Table 5), which play a different role. Articles with higher ChatGPT scores or more citations were more likely to be analytical in the sense of including terms in their abstracts that suggest analytical thinking.

The same is true for abstracts with importance or novelty claims that situate the study in a wider context. These all seem indicative of either higher quality research or at least an abstract that suggests or claims higher quality research, so the higher citation rates and ChatGPT scores seem reasonable for these themes.

Research with abstracts mentioning data details or evidence tends to get more citations and higher ChatGPT scores than articles that do not mention them in abstracts. The remaining articles tended to also use data, but without focusing on it in the abstract. Thus, this theme may be a side-effect of a more analytical abstract. In contrast, it is perhaps unsurprising that abstracts reporting the use of complex statistics are more cited and higher scoring, and it seems likely that they tend to be higher quality because the authors have gone beyond simple statistics to investigate more complex relationships. Two qualitative methods (ethnography and grounded theory) are apparently favored by ChatGPT without being more cited. Including explicit descriptions of theory also associates with both higher ChatGPT scores and more citations.

Table 4. WATA themes, other than topics, derived from terms occurring disproportionately often in high ChatGPT scoring or highly cited tourism and leisure articles 2011-2020.

Group and theme	High ChatGPT terms*	High citation terms*
Analytical: analysis or evaluation.	Critical, complexities, debate, simultaneously (34 terms)	Furthermore, **relationship, between, role**, **effect** (8 terms)
Framing: Explaining study importance or novelty.	Advance, increasingly, critical, novel	Managerial [implications], [recommendations for] future [research], confirm [hypothesis]
Framing: Situating study in a wider context or in relation to prior literature.	Address, highlight, [research] gap, context, discourse, outcome (18 terms)	**Literature**, study
Methods: Data details or evidence	Experimental, experiment, **empirical**, **empirically**, tested, evidence	Collected, sample, tested, **empirical**, **empirically**
Methods: Complex statistics such as SEM or mediating relationships	Mediate, multilevel, structural, **mediated**, interaction, **moderating**, indirect, **modelling** (12 terms)	Mediator, **mediated**, hypothesized, moderator, **moderating**, partial, causal, structural, **modelling** (17 terms)
Methods: Description	Through, approach	
Methods: Mixed methods	Mixed-methods	
Methods: Ethnography	Ethnographic, fieldwork	
Methods: Grounded theory	Grounded	
Methods: Hypothesis test		Hypotheses, test
Methods: Behavioral intention analyses		Influence, antecedent, behavior, behavioral, behavioral, behavior, intention, perceived
Style: I/We argue/find that/demonstrate	I, we, argue, find, demonstrate	
Style: Structured abstract headings		Findings, implications
Theory: A theoretical framework or model proposed or used	Drawing, logic, **framework**, **construct, proposed, propose, theoretical**, **model, theory**, conceptualization (16 terms)	**Framework, construct, proposed, propose, theoretical, model, theory**
Theory: Neoliberalism	Neoliberal	
Theory: the mobilities paradigm	Mobility, mobilities	
Theory: Theory of planned behavior		TPB

**Square brackets are used to indicate the typical context of a term.*

Many tourism and leisure topics tend to be more cited, higher scoring, or both, including tourism itself (Table 5). It seems reasonable to suspect that research on topics that are more cited and attract higher ChatGPT scores tends to be generally higher quality. This might be because it is about an issue that is widely considered important, such as the environment, because it associates with a finer-grained analysis (e.g., consumers, emotions), or because it takes a more analytical approach (e.g., co-creation, authenticity).

Topics favored by ChatGPT but not citations (at least at the level of statistical significance used here) seem to be mainly associated with more in-depth analyses (e.g., gender, identities, embodied, encounters). In contrast, the more cited but not higher scoring topics also include some that seem to be associated with (then) emerging issues, such as Covid-19 and peer-to-peer accommodation, where early research may have been original but less robust than mature research, attracting many citations as the topic expanded. Since ChatGPT does not seem to consider publication years, it is possible that it gives insufficient credit to early research on a topic. Conversely, this early research might have "too many" citations for its quality due to being an early contribution to an expanding topic.

Table 5. WATA topic-based themes derived from terms occurring disproportionately often in high ChatGPT scoring or highly cited tourism and leisure articles 2011-2020.

Topic theme	High ChatGPT terms	High citation terms
Environment	**Climate, pro-environmental**	**Climate, pro-environmental**, green, environmentally
Co-creation of tourism experiences	**Co-creation**	**Co-creation**
Authenticity within tourism	**Authenticity**	**Authenticity**
Psychological considerations	**Psychological**	**Psychological**
Emotional reactions	**Emotions, emotional**	**Emotions, emotional**, affective
Tourism destinations	**Tourism**	**Tourism**, tourist, destination
Consumer perspective: consumers, marketing, online, purchase intentions	**Consumer**, consumption	**Consumer**, attitude, trust, hedonic
Business performance	Firm	
Embodied experiences	Embodied	
Gender	Gendered	
Personal identities	Identities, identity	
Local communities	Indigenous, communities	
Political considerations	Political, politics	
Service encounters	Encounter	
Peer-to-peer accommodation		AirBnb, peer-to-peer
Covid-19		Covid-19, pandemic
CSR: Corporate Social Responsibility		CSR, responsibility
Customer services		Customer, service, satisfaction, loyalty, revisit
Frontline employees		Frontline
Hospitality industry		Hospitality, employee, hotel, guest, restaurant
Innovation		Innovativeness
Memorable tourism experiences		Memorable
User-generated content		User-generated
Traveler		Travel
Word-of-mouth		Word-of-mouth, ewom

RQ3a: Themes for lower ChatGPT research quality scores and/or less cited articles

Many themes were also found for lower scoring and/or less cited research. Surprisingly, some research framing themes associated with lower ChatGPT scores and one also associated with fewer citations (Table 6). This might be a second order style effect, with higher quality research tending to be phrased differently either because of journal style or author preferences. In terms of methods, some attracted low ChatGPT scores, and others were comparatively rarely cited. Some specific qualitative methods seemed to be less cited, perhaps because fewer citations are common for qualitative research, or a higher proportion of qualitative research papers are published as book chapters or books and not indexed by Scopus. In contrast, ChatGPT tended to give low scores to articles with basic statistics and to surveys, especially when using convenience sampling.

Table 6. WATA themes, other than topics, derived from terms occurring disproportionately often in low ChatGPT scoring or less cited tourism and leisure articles 2011-2020.

Group and theme	Low ChatGPT terms	Low citation terms*
Framing: objectives/problem	**Objective, problem, purpose**	**Objective, problem, purpose**, solution, question
Framing: show/determine	Determine, showed	
Framing: research recommendations	Recommendations	
Methods: Descriptive statistics	Descriptive	
Methods: Factor analysis	Factor, exploratory	
Methods: Statistics	Statistical, significant, t-test, difference, SPSS	
Methods: Surveys	Collected, questionnaire, respondent, conducted, convenience, sampling, survey, surveyed	
Methods: interviews, focus groups		Group, interview
Methods: Case study or case		Case
Methods: Contingent valuation method		Valuation
Style: Vague or wordy	According, main	[the/in] fact [that], [carried] out
Style: Past tense	Were, was, had, indicated, used, aimed	Became
Style: future tense	Will	Could
Style: indirect		[this] article, [the] author
Style: Roman numerals		I

**Square brackets are used to indicate the typical context of a term.*

Several broad topics or aspects of topics are less cited and tend to receive lower ChatGPT scores (Table 7). For educational research, this might be due to some educational research being conducted by interested tourism and leisure researchers that also teach, rather than education research specialists. For research mentioning individual countries, the reason might be that the study is more of local relevance, hence lowering its ChatGPT score (due to weak significance) and may attract fewer citations since fewer articles would cover the same country. It is not clear why the other topics attract lower scores and fewer citations. Several topics attracted only low ChatGPT scores and others attracted few citations. Any topic might attract few citations if it is a small specialty, a low citation area or an area in which many citations come from sources not indexed in Scopus, such as national journals, books and book chapters. This might also be the reason why tourism research tends to be more cited whereas leisure research tends to be less cited.

Table 7. WATA topic-based themes derived from terms occurring disproportionately often in low ChatGPT scoring or less cited tourism articles 2011-2020.

Topic theme	Low ChatGPT terms	Low citation terms
Education	Careers, **student, education, undergraduate, university**, study, graduate, **teaching, course**, curriculum, faculty, educator, training, **college, program**	**Student, education, undergraduate, university**, **teaching, course**, **college, program**, school, educational, programme
Individual countries	**Kazakhstan, Republic, Russian, Russia**, Indonesia, Malaysia, Turkey	**Kazakhstan, Republic, Russian, Russia**, national, country, states, ministry, federation
Employees	Staff, job	Working
Government policy	**Government**	**Government**
Rural tourism	**Agricultural**	**Agricultural**
Temporal trends or changes	**Year**, trend	**Year**
Overview and review of research	Overview, published [research]	
Demographic variables	Age, demographics	
Customer satisfaction	Quality, satisfaction, satisfied	
Tourism industry	Industry	
Facilities for visitors	Facilities	
Food (in restaurants)	Food	
Amateur arts/sports performances		Amateur
Culture of tourists or destination		Culture
Economic growth or development		Economic
Events		Event
Forests as tourist sites		Forest
Venue		Venue
Cultural heritage		History, century, historical, cultural
Leisure or recreation activities		Leisure, recreation, activity, activities, recreational
Modern		Modern
Region within a country		Territory, territories, district
Sports		Sport, sporting
Art		Art
Wider society		Society
Urban tourism		City, urban

5. Discussion

In answer to RQ1, the normalized citation rate of a tourism and leisure article correlates positively and moderately (r=0.288) with its ChatGPT research quality score. This aligns with a similar positive correlation previously found for the related discipline of business between ChatGPT and citation rates and between ChatGPT and research quality scores

(Thelwall & Yaghi, 2025). The average research quality score for articles (ChatGPTn = 2.628) and journals (ChatGPTn = 2.592) on the four-point scale is lower than the Research Excellence Framework 2021 evaluation results for the overall field of 'Sport and Exercise Sciences, Leisure and Tourism', where the average score for outputs was 3.111 (REF, 2021b).

Although some progress has been achieved, tourism research is still lacks theoretical rigor overall (Nunkoo & Armbrecth, 2025). Most studies have not related concepts to a broader theoretical structure and the field has failed to develop tourism-specific theories (McCabe, 2024). At the methodological level, there are problematic uses of qualitative research approaches (Nunkoo, Smith, & Ramkissoon, 2013; Riley and Love, 2000), statistical techniques (Nunkoo, Ramkissoon, & Gursoy, 2013), and survey design (Dolnicar, 2020) that pose challenges for research rigor. Disciplinary parochialism, repeated re-conceptualizations of, and lack of consensus over dominant paradigms such as sustainable tourism, and case-study based research have also hindered the field's intellectual progress (Moyle, Moyle, Ruhanen, Weaver, & Hadinejad, 2021).

In terms of research impact, tourism researchers have for a long time now debated the enduring gap between tourism research and tourism practice (Jones & Walmsley, 2022; Ritchie & Ritchie, 2002). In support of our findings, Thomas and Ormerod (2017) empirically demonstrate that the non-academic impact of tourism is low, while studies on the performance of tourism research in UK Research Excellence Framework assessment suggest that although the field has attained a good scholarly standard, engagement with the public and businesses remains negligible (Phillips et al., 2020; Tribe & Paddison, 2024). Tourism has failed to contribute much to broader development agendas such as the United Nations Sustainable Development Goals (Rasoolimanesh, Ramakrishna, Hall, Esfandiar, & Seyfi, 2023). These arguments provide plausible explanations for the low ChatGPT research quality scores revealed by this study.

For RQ2, the average citation rate of relatively large tourism and leisure journals correlates positively and strongly (r=0.598) with their mean ChatGPT research quality score. Although based on a different approach, Thomas and Ormerod (2017) found a significant relationship between the scholarly impact of tourism research and its non-academic impact: top-cited articles were more likely to be used by policy makers. While LLMs have not been used previously to rank journals in tourism and allied fields, the journals' performance based on the MNLCS, which differs from that of the ChatGPTn scores (see Appendix 1), confirms the leadership of *Tourism Management*, *Journal of Travel Research*, *Annals of Tourism Research*, *International Journal of Hospitality Management*, and *Journal of Sustainable Tourism* (*JoST*) established by previous rankings (e.g., Gursoy & Sandstrom, 2016).

The difference in the journals' performance in the two rankings is expected because unlike the MNLCS, the ChatGPTn include broader considerations such as their significance. As Jamal, Smith and Watson (2008) note, journals are likely to differ in their performance not only based on their scholarly impact, but also on their scope, influence, degree of specialization, and the disciplines and sub-disciplines they serve. The *JoST*'s leading position in the ChatGPTn ranking, although the score differences between the journals are relatively narrow, (see Appendix 1), can be attributed to its specialized focus on sustainable tourism which remains a dominant discourse, more than other concerns, in governments' policy narratives (Bramwell, 2011), providing more collaboration opportunities between

tourism scholars and policymakers to produce impactful research (Font, Higham, Miller, & Pourfakhimi, 2019).

The performance of leisure journals also warrants attention. Their relatively high ChatGPT scores for their low citation rates (Figure 6) suggest that while they are cited less than tourism journals, possibly because of the relatively smaller size of the field compared to tourism, in terms of research quality, leisure research perform as well as tourism. For example, there is a negligible difference in the ChatGPT scores of *Leisure Studies* and *JoST*, *TM*, and *ATR*. Leisure studies have reached a certain level of conceptual, theoretical, and methodological maturity (Henderson, Presley, & Bialeschki, 2004) and have been impactful on people's lives by focusing, for example, on 'micro spaces' such as gardens and homes.

The ageing population in several countries has placed leisure activities at the center of active and healthy ageing policies, opening exciting opportunities for dialogues between leisure scholars, practitioners, and policymakers (Nimrod, Janke, & Kleiber, 2016). *Tourist Studies* also perform well in our ranking, possibility because of its humanistic-oriented research focusing on tourism as a social phenomenon. Furthermore, a closer look at the highest scoring article (GPTn=3.87) by Benali and Ren's (2019) published in the journal describes the researchers' lived experience with the local community over four weeks of field work as a volunteer tourist at a Nepalese orphanage. In addition to the theoretical and methodological rigor of the study, the authors' use of pictures provides evidence of their deep engagement with the locals achieved by helping them deal with the problem of head lice to improve their hygienic conditions.

The WATA analysis suggests interesting nuances between quantitative and qualitative techniques in tourism studies, as well as within both quantitative and qualitative techniques. In a recent study on the typology of quantitative approaches to discovery, Dolnicar, Zinn and Demeter (2024) emphasize on the need to distinguish between various approaches to quantitative research in tourism given their implications for internal and external validity. In line with their arguments, this study finds that quantitative studies represented by terms such as 'experiment', and 'empirical', are associated with both high citations and ChatGPT scores, suggesting higher research quality, while those based on surveys and convenience sampling are associated with low ChatGPT scores. Supporting our findings, there are known problems with survey-based research (Dolinar, 2018, 2020) while the value of experimental designs in enhancing theoretical and methodological rigor of tourism research has also been established (Dolnicar et al., 2024). High ChatGPT scoring or highly cited tourism articles also make use of advanced statistical techniques. The prevalence of advanced statistical techniques such as structural equation modeling has increased in tourism and hospitality research (do Valle, & Assaker, 2016; Nunkoo et al., 2013) and have allowed researchers to study real-life phenomenon by testing multiple hypotheses simultaneously that approximate the multidimensional nature of reality (Bagozzi & Yi, 2012).

Qualitative research represented by terms such as 'ethnography' and 'grounded theory' are favored by ChatGPT without being well-cited. Although, traditionally, qualitative studies have struggled for legitimacy and rigor in tourism (Jamal & Hollishead, 2001) and suffered from the beliefs that their findings are difficult to translate into tourism policy (Riley & Love, 2000), the WATA high ChatGPT terms are suggestive of their high research quality. Not only have qualitative studies progressed in terms of conceptual and methodological rigor (McGinley, Wei, Zhang, & Zheng, 2021), but they have also reclaimed their relevance to practice (Allen, Kelly, & Hatala, 2024; Ross, 2022). Their low citation potential, however,

can be explained by the dominance of positivistic paradigm over more subjective approaches across several research topics in tourism and hospitality.

The writing style also influences research quality and citation potential of an article. Tribe (2018) emphasizes the clarity of expression, engaging writing, a well-written abstract as important aspects of research rigor. The WATA analysis suggests that studies that frame the research problem in a wider context while considering prior literature and theories, compared to a more descriptive and narrow framing or those that emphasize a specific setting (e.g., a country), associate with both higher ChatGPT score and higher citation. These studies are able to develop compelling research questions that have the potential "to create or change scholarly consensus in the relevant audience and prior research" (Dorobantu, Gruber, Ravasi, & Wellman, 2024), generating substantive new tourism knowledge (Dolnicar, 2020).

Theoretically informed studies also score highly by ChatGPT and have high citation potential because they go beyond simply describing a tourism phenomenon by explaining how, why, and under what circumstances it occurs while providing important implications for tourism practice (Nunkoo & Armbrecth, 2025). Studies using TPB are highly cited given their popularity in tourism and hospitality research (Ulker-Demirel & Ciftci, 2020), but this does not necessarily mean that these studies are of high quality. Ulker-Demirel and Ciftci's (2020) review of the TPB found that several studies used the theory as an explanation for the concept without further explanation or application while others tested the theory without building on it. Furthermore, suggesting that intention is a good approximation of behavior which is the case in several TPB-based studies is misleading given the weak association between behavioral intention and actual behavior. Such imprecise language use is problematic for research quality (Greene & Dolnicar, 2024).

In comparison to a previous analysis of the value of ChatGPT for research assessment for the library and information science field, the statistical results, where they overlap, are not substantially different. The most interesting comparison is for the high and low scoring topics and methods, with an important difference being that ChatGPT tended to give higher scores to library and information science abstracts with explicit novelty and research context claims (Thelwall, 2025b). Both novelty (in the framing theme) and aspects of context were evident in the leisure and tourism results but were much more prominent for the library and information science field.

5.1 Practical Implementation process

If ChatGPT is used to support a research evaluation of journal articles, then the following steps are recommended. First, identify a benchmark collection of at least 1000 journal articles. This should be matched to the years of the articles to be evaluated. A benchmark collection is essential because ChatGPT does not use the same scale as human experts, and the collection will allow this issue to be factored out. The benchmark collection could be ignored if the evaluation is comparing multiple large sets of articles (e.g., from a set of departments) since the sets will serve as benchmarks for each other.

Second, decide on the definition of research quality to be used and translate this into prompts. If the REF definitions are used, then these can be recycled from the current paper. If other definitions are used then they should be expressed in natural language and, especially if they are radically different from the REF guidelines, testing should be conducted (as in the current paper) to check whether they give plausible results.

Third, submit the articles to be evaluated and the benchmark set to ChatGPT 4o-mini (or a successor, such as ChatGPT 5o-mini, or another LLM with equivalent power) with the quality definitions as system prompts and the simple user prompt, "Score this journal article:" followed by the article title, the text "\nAbstract\n" and the abstract. This should be submitted at least five times. Increase this to 30 if higher precision is needed, such as if there are few articles to evaluate. The submission can be achieved automatically with the ChatGPT API (see above for the code) and currently is cheapest in "batch mode". The submission file should be ordered randomly and submitted five (or 30) times rather than including five (or 30) identical requests in the same file, to minimize the chance of one score influencing another for the same article. If the web interface of ChatGPT is used instead of the API then a new session should be started for each score because scores interfere with each other in a single web session. This problem does not occur in the API when requests are sent as separate calls, as above.

Fourth, extract the scores from the ChatGPT reports using Webometric Analyst (in two steps, first extracting the reports from the API Json and then extracting the scores from the reports. The average of all the scores should be used as the overall ChatGPT score for the article.

Fifth, convert the overall ChatGPT scores for each article into usable information in one of three ways. The simplest it to rank the scores and assign a percentile against the benchmark collection or, if no benchmark is used, against the whole set. For example, article A with a ChatGPT score of 3.1* might have a percentile of 96%, meaning that this score is at least as high as 96% of articles in the collection. Alternatively, the ChatGPT scores can be converted into exact scores (e.g., REF star scores) through a look-up table. This table could be constructed either by identifying the thresholds with a benchmark set or by deciding the percentage of articles that should be in each class (e.g., based on a previous exercise) and selecting the thresholds to achieve these percentages. Alternatively, a formula might be used to assign partial scores (e.g., 2.4*), perhaps by starting with the lookup table thresholds and designing a function to interpolate between the values. Using simple percentiles would be simpler and more straightforward if fine grained scores are needed.

Sixth, use the results to make the evaluation. If individual articles are being evaluated then the expert (or not so expert) evaluators should be provided with the percentiles to help their judgements (this is how citation data was used in REF2021). For example, if the expert is unsure whether an article should be assigned a 3* or 4* but the ChatGPT score is only in the top 50% of scores then they might decide on 3*. Alternatively, if two or more sets of articles are to be evaluated (e.g., to compare departments or journals) then the average percentile could be calculated for each one. For this purpose, average ChatGPT scores would be statistically equivalent, but the results would be less intuitive because the ChatGPT scores are not a useful scale.

6. Conclusion

This study is a modest attempt to evaluate the quality of tourism research by going beyond citation-based indicators as measures of research quality based on the UK Research Excellence Framework's guidelines that reflect the common Global North definition of research quality (Langfeldt et al., 2020). It addresses the concerns of some scholars with the use of citation and productivity metrices to evaluate tourism research. An article may be highly cited simply because it deals with a hot or popular topic, but this does not necessarily account for its research quality. On the other hand, a study may be socially impactful, but it is cited less because of its niche focus (Jamal et al., 2008). Tourism studies' ultimate goals

should lead to new discoveries and create positive changes in societies (Benjamin et al., 2024; Dolnicar, 2025) which imply rigor, originality, and significance in their conceptualization and execution. Although measuring the originality and real-world impact of research is challenging, this study leverages the benefits of large language models like ChatGPT to develop and start scholarly discussions on a new form of research evaluation.

The themes associated with high and low ChatGPT scores and citation rates might be useful for tourism and leisure researchers to reflect about what constitutes high quality research in the field, for example in terms of mixed-methods approaches or experiment-based studies or the incorporation of theory or those that inspire changes in societies. However, these themes should not be over-interpreted as *proving* that the associated research tends to be good or bad. Tourism as well as broader literature provides well established guidelines to develop original and theoretically and methodologically robust, and impactful studies that researchers are encouraged to follow (e.g., Dorobantu et al., Greene & Dolnicar, 2024; Tribe & Paddison, 2024) to develop high quality research.

The finding that mean normalized ChatGPT scores tend to correlate positively with mean normalized citation rates for tourism and leisure journals tends to mutually support the value of the rankings from both. This is because they seem to be independent sources and for both citations (Thelwall et al., 2023) and ChatGPT (Thelwall, & Yaghi, 2025), there is evidence that they partly reflect research quality. In this context, the advantage of ChatGPT is that it can score articles from the current year, whereas citations take time to accrue, so ChatGPT-based journal rankings might be more sensitive to recent changes in journals, such as editorial board changes or direction shifts, as well as being available for new journals in advance of any citation-based indicators for them. In this context, the ChatGPT scores for any evaluation will need to be scale-transformed, such as by a look-up table or mathematical formula to adjust the scale to one that would be more typical for humans (e.g., using the full range of 1* to 4*, if a Research Excellence Framework scale is to be used). Alternatively, simple rankings of the results avoids the need for any kind of scaling. Given the relatively weak correlations found here and the moderate correlations with research quality for the wider REF field, the ChatGPT scores are far from quality guarantees. Instead they, like citations, should be used as pointers: much less valuable than expert judgment but useful when expertise is unavailable because there are too many articles to evaluate, no experts available, or the available expertise does not match the topic of some or all of the research assessed.

Since citation-based indicators seem to be deeply embedded into academic perceptions of research quality in tourism, it seems unlikely that LLM-based rankings will replace them in the near future, but they could be used as a second opinion, especially for authors that are skeptical about existing tourism and leisure rankings. Of course, journal-based rankings should not replace expert review of individual journal articles for important research evaluation goals (sfdora.org). Moreover, if LLM-based rankings become systematically used, then authors and journal editors might start to "cheat" by tailoring abstracts for higher ChatGPT scores, which might degrade their value and undermine the main communication purposes of abstracts. The effectiveness of such approaches therefore needs to be tested and strategies developed to mitigate against them.

This study is limited by the journal selection used, which delineates the topic. Different results may have been obtained from a more inclusive or less inclusive set. The citation data is also restricted to documents indexed by Scopus, which has biases against non-English publications (Mongeon & Paul-Hus, 2016). The grouping of terms into WATA themes is

subjective and alternative plausible themes may have been obtained for the same terms. In addition, the terms analyzed are based on a relatively arbitrary statistical threshold. Finally, whilst the computer architecture of LLMs is known, the pathway that ChatGPT 4o-mini follows for any of its research quality reports is opaque, so it is impossible to be certain about any cause-and-effect claim or whether other LLMs, including future versions of ChatGPT are likely to perform substantially differently. In particular, it is not known whether ChatGPT can identify citation data for the articles submitted, for example by associating the title and abstract with an article and having access to citation data for articles. Moreover, ChatGPT is too reluctant to give high (4*) or low (1*) scores so its raw output may need norm referencing to be meaningful.

References

Ali, F., Rasoolimanesh, S. M., Sarstedt, M., Ringle, C. M., & Ryu, K. (2018). An assessment of the use of partial least squares structural equation modeling (PLS-SEM) in hospitality research. *International journal of contemporary hospitality management*, *30*(1), 514-538.

Allen, L. P., Kelly, C., & Hatala, A. R. (2024). Answering tough questions: Why is qualitative research essential for public health?. *Australian and New Zealand Journal of Public Health*, *48*(3), 100157.

Bagozzi, R. P., & Yi, Y. (2012). Specification, evaluation, and interpretation of structural equation models. *Journal of the academy of marketing science*, *40*, 8-34.

Bartneck, C., & Kokkelmans, S. (2011). Detecting h-index manipulation through self-citation analysis. *Scientometrics*, *87*(1), 85-98.

Benali, A., & Ren, C. (2019). Lice work: Non-human trajectories in volunteer tourism. Tourist Studies, 19(2), 238-257.

Benckendorff, P., & Zehrer, A. (2013). A network analysis of tourism research. *Annals of tourism research*, *43*, 121-149.

Benjamin, S., Lee, K. S., & Boluk, K. (2024). Shit has to change, right? A call for "good trouble" in tourism. *Journal of Travel Research*, 00472875241276542.

Bramwell, B. (2011). Governance, the state and sustainable tourism: A political economy approach. *Journal of sustainable tourism*, *19*(4-5), 459-477.

Brauer, R., Dymitrow, M., & Tribe, J. (2019). The impact of tourism research. *Annals of Tourism Research*, *77*, 64-78.

Buckley, R. (2019). Tourism publications as newly tradeable commodities: Academic performance, prestige, power, competition, constraints and consents. *Annals of Tourism Research*, *74*, 121-133.

Buckley, R. (2023). Originality in research publication: Measure, concept, or skill?. *Journal of Travel Research*, *62*(5), 1159-1163.

Czernek-Marszałek, K., & McCabe, S. (2024). Sampling in qualitative interview research: criteria, considerations and guidelines for success. *Annals of Tourism Research*, *104*, 103711.

do Valle, P. O., & Assaker, G. (2016). Using partial least squares structural equation modeling in tourism research: A review of past research and recommendations for future applications. *Journal of Travel Research*, *55*(6), 695-708.

Dolnicar, S. (2018). A reflection on survey research in hospitality. *International Journal of Contemporary Hospitality Management*, *30*(11), 3412-3422.

Dolnicar, S. (2020). Survey research in tourism: a perspective paper. *Tourism Review*, *75*(1), 20-23.

Dolnicar, S. (2025). Not enjoying the publish or perish culture? You have two options only: Fuel it or resist it. Which will you choose?. *Annals of Tourism Research*, *110*.

Dolnicar, S., Zinn, A. K., & Demeter, C. (2024). A typology of quantitative approaches to discovery. *Annals of Tourism Research*, *104*, 103704.

Dorobantu, S., Gruber, M., Ravasi, D., & Wellman, N. (2024). The AMJ management research canvas: A tool for conducting and reporting empirical research. *Academy of Management Journal*, *67*(5), 1163-1174.

Font, X., Higham, J., Miller, G., & Pourfakhimi, S. (2019). Research engagement, impact and sustainable tourism. *Journal of sustainable tourism*, *27*(1), 1-11.

Freedman, D., & Diaconis, P. (1981). On the histogram as a density estimator: L 2 theory. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete, 57(4), 453-476. https://doi.org/10.1007/BF01025868

Greene, D., & Dolnicar, S. (2024). On the importance of precise language use. *Annals of Tourism Research*, *104*, 103707.
Gursoy, D., & Sandstrom, J. K. (2016). An updated ranking of hospitality and tourism journals. *Journal of Hospitality & Tourism Research*, *40*(1), 3-18.
Hall, C. M. (2011). Publish and perish? Bibliometric analysis, journal ranking and the assessment of research quality in tourism. *Tourism management*, *32*(1), 16-27.
Henderson, K. A., Presley, J., & Bialeschki, M. D. (2004). Theory in recreation and leisure research: Reflections from the editors. *Leisure sciences*, *26*(4), 411-425.
Ioannidis, J. (2024). Features and signals in precocious citation impact: a meta-research study. *bioRxiv*, 2024-10.
Jamal, T., & Hollinshead, K. (2001). Tourism and the forbidden zone: The underserved power of qualitative inquiry. *Tourism management*, *22*(1), 63-82.
Jamal, T., Smith, B., & Watson, E. (2008). Ranking, rating and scoring of tourism journals: Interdisciplinary challenges and innovations. *Tourism Management*, *29*(1), 66-78.
Jones, C. R., & Walmsley, A. (2022). A change would do you good: advances in research impact in sustainable tourism and some 'home truths' for the sector. *Journal of Sustainable Tourism*, *30*(9), 2073-2088.
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.
Latona, G. R., Ribeiro, M. H., Davidson, T. R., Veselovsky, V., & West, R. (2024). The AI Review Lottery: Widespread AI-Assisted Peer Reviews Boost Paper Scores and Acceptance Rates. arXiv preprint arXiv:2405.02150.
Law, R., & Chon, K. (2007). Evaluating research performance in tourism and hospitality: The perspective of university program heads. *Tourism Management*, *28*(5), 1203-1211.
Lee, K. S., & Benjamin, S. (2023). The death of tourism scholarship… unless…. *Annals of Tourism Research*, *98*, 103520.
Leung, X. Y., Fong, L. H. N., Xue, X., & Mattila, A. S. (2024). What makes experimental research publishable in leading hospitality and tourism journals? Perspectives of editorial board members. *International Journal of Contemporary Hospitality Management*, *36*(4), 1418-1431.
Li, J., & Qiu, X. (2024). Research on innovation quality measurement of papers based on large language model. Information Studies: Theory & Application, 53(4), 345-356. https://doi.org/10.12345/abc123
McCabe, S. (2024). Theory in tourism. *Annals of Tourism Research*, *104*, 103721.
McGinley, S., Wei, W., Zhang, L., & Zheng, Y. (2021). The state of qualitative research in hospitality: A 5-year review 2014 to 2019. *Cornell Hospitality Quarterly*, *62*(1), 8-20.
Means, W. T., & Mowatt, R. A. (2024). Philosophy of science and leisure research: an exploratory analysis of research paradigms. *Leisure/Loisir*, *48*(1), 123-147.
Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: a comparative analysis. Scientometrics, 106, 213-228.
Moyle, B., Moyle, C. L., Ruhanen, L., Weaver, D., & Hadinejad, A. (2020). Are we really progressing sustainable tourism research? A bibliometric analysis. *Journal of Sustainable Tourism*, *29*(1), 106-122.
Nimrod, G., Janke, M. C., & Kleiber, D. A. (2016). Leisure and aging qualitative research 15 years into the third millennium. *Journal of Leisure Research*, *48*(1), 12-14.
Nunkoo, R., & Armbrecht, J. (2025). What theory is and is not? The need for theorizing in tourism research. *Tourism Management*, *109*, 105150.

Nunkoo, R., Hall, C. M., Rughoobur-Seetah, S., & Teeroovengadum, V. (2019). Citation practices in tourism research: Toward a gender conscientious engagement. *Annals of Tourism Research*, *79*, 102755.

Nunkoo, R., Ramkissoon, H., & Gursoy, D. (2013). Use of structural equation modeling in tourism research: Past, present, and future. *Journal of Travel Research*, *52*(6), 759-771.

Nunkoo, R., Smith, S. L., & Ramkissoon, H. (2013). Residents' attitudes to tourism: A longitudinal study of 140 articles from 1984 to 2010. *Journal of Sustainable Tourism*, *21*(1), 5-25.

Nunkoo, R., Thelwall, M., Ladsawut, J., & Goolaup, S. (2020). Three decades of tourism scholarship: Gender, collaboration and research methods. *Tourism management*, *78*, 104056.

Olaru, J. M. D., & Hofacker, C. F. (2009). Rigor in tourism research: Formative and reflective constructs. *Annals of Tourism Research*, *36*(4), 730-734.

Phillips, P. A., Page, S. J., & Sebu, J. (2020). Achieving research impact in tourism: Modelling and evaluating outcomes from the UKs Research Excellence Framework. *Tourism Management*, *78*, 104072.

Polat, E., Çelik, F., Arici, H. E., & Köseoglu, M. A. (2024). Predictors of citations: an analysis of highly-cited-papers in hospitality and tourism research using a machine learning approach. *Current Issues in Tourism*, 1-22.

Rasoolimanesh, S. M., Ramakrishna, S., Hall, C. M., Esfandiar, K., & Seyfi, S. (2023). A systematic scoping review of sustainable tourism indicators in relation to the sustainable development goals. *Journal of sustainable tourism*, *31*(7), 1497-1517.

REF (2021a). Panel criteria and working methods (2019/02). Retrieved from https://2021.ref.ac.uk/publications-and-reports/panel-criteria-and-working-methods-201902/index.html (access date: 15 Feb, 2025).

REF (2021b). Results and submissions: introduction to the REF results. https://results2021.ref.ac.uk/unit-of-assessment-summary/24

Riley, R. W., & Love, L. L. (2000). The state of qualitative tourism research. *Annals of tourism research*, *27*(1), 164-187.

Ritchie, R. J., & Ritchie, J. B. (2002). A framework for an industry supported destination marketing information system. *Tourism Management*, *23*(5), 439-454.

Rodriguez Sanchez, I., Mantecón, A., Williams, A. M., Makkonen, T., & Kim, Y. R. (2022). Originality: the holy grail of tourism research. *Journal of Travel research*, *61*(6), 1219-1232.

Ross, K. (2022). Reclaiming impact in qualitative research. In *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research* (Vol. 23, No. 2). DEU.

Ryan, C. (2005). The ranking and rating of academics and journals in tourism research. *Tourism Management*, *26*(5), 657-662.

Sánchez, I. R., Makkonen, T., & Williams, A. M. (2019). Peer review assessment of originality in tourism journals: critical perspective of key gatekeepers. *Annals of Tourism Research*, *77*, 1-11.

Sheather, S. J., & Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. Journal of the Royal Statistical Society: Series B (Methodological), 53(3), 683-690. https://doi.org/10.1111/j.2517-6161.1991.tb01857.x

Thelwall, M. (2017). Three practical field normalised alternative indicator formulae for research evaluation. Journal of informetrics, 11(1), 128-151.

Thelwall, M. (2021). Word association thematic analysis: A social media text exploration strategy. San Rafael, CA: Morgan & Claypool.

Thelwall, M. (2023). Word Association Thematic Analysis: Insight Discovery from the Social Web. SN Computer Science, 4(6), 827.

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. (2025a). 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. (2025b). Can ChatGPT replace citations for quality evaluation of academic articles and journals? Empirical evidence from library and information science. *Journal of Documentation*, 81(4), 1078–1094. https://doi.org/10.1108/JD-03-2025-0075

Thelwall, M., & Nevill, T. (2021). Is research with qualitative data more prevalent and impactful now? Interviews, case studies, focus groups and ethnographies. *Library & Information Science Research*, *43*(2), 101094.

Thelwall, M., & Yaghi, A. (2025). In which fields can ChatGPT detect journal article quality? An evaluation of REF2021 results. Trends in Information Management, 13(1), 1-29. https://arxiv.org/abs/2409.16695

Thelwall, M., Jiang, X., & Bath, P. A. (2024). Evaluating the quality of published medical research with ChatGPT. arXiv preprint arXiv:2411.01952.

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

Thomas, R. (2024). A perspective on official research performance evaluation in tourism. In C.Cooper and C.M. Hall (Eds.), *How to Get Published in the Best Tourism Journals* (pp. 189-201). Edward Elgar Publishing.

Tribe, J. (2018). Creating and curating tourism knowledge. *Annals of Tourism Research*, *73*, 14-25.

Tribe, J., & Paddison, B. (2024). Paths from knowledge and theory development to impact. *Annals of Tourism Research*, *104*, 103687.

Ulker-Demirel, E., & Ciftci, G. (2020). A systematic literature review of the theory of planned behavior in tourism, leisure and hospitality management research. *Journal of Hospitality and Tourism Management*, *43*, 209-219.

Usakli, A., & Kucukergin, K. G. (2018). Using partial least squares structural equation modeling in hospitality and tourism: do researchers follow practical guidelines?. *International Journal of Contemporary Hospitality Management*, *30*(11), 3462-3512.

Viana-Lora, A. (2023). The societal impact of tourism research of the Research Excellence Framework 2021. *Journal of Policy Research in Tourism, Leisure and Events*, 1-16.

Viana-Lora, A., Nel-lo-Andreu, M. G., & Anton-Clavé, S. (2023). Advancing a framework for social impact assessment of tourism research. *Tourism and Hospitality Research*, *23*(4), 494-505.

Walle, A. H. (1997). Quantitative versus qualitative tourism research. *Annals of tourism research*, *24*(3), 524-536.

Wang, J. (2013). Citation time window choice for research impact evaluation. Scientometrics, 94(3), 851-872.

Appendix 1: Journal average citation rates (MNLCS) and average normalized ChatGPT scores 2011-2020 (ranking is based on the average ChatGPT score)

Journal	Average ChatGPTn (Rank)	MNLCS (Rank)	Articles
Journal of Sustainable Tourism	2.901 (1)	1.199 (5)	837
Tourist Studies	2.893 (2)	0.885 (29)	192
Tourism Geographies	2.828 (3)	1.091 (12)	382
Annals of Tourism Research	2.825 (4)	1.260 (3)	740
Journal of Travel Research	2.825 (4)	1.261 (2)	621
Journal of Outdoor Recreation and Tourism	2.788 (6)	0.935 (22)	245
Tourism Management	2.763 (7)	1.330 (1)	1584
Leisure Studies	2.757 (8)	0.887 (27)	425
International J Contemporary Hospitality Management	2.741 (9)	1.156 (7)	1125
Journal of Destination Marketing and Management	2.737 (10)	1.189 (6)	385
International Journal of Hospitality Management	2.733 (11)	1.222 (4)	1392
Cornell Hospitality Quarterly	2.727 (12)	0.976 (18)	302
Journal of Hospitality and Tourism Research	2.723 (13)	1.089 (14)	334
Leisure Sciences	2.719 (14)	0.929 (23)	257
Current Issues in Tourism	2.682 (15)	1.107 (11)	689
Journal of Tourism and Cultural Change	2.675 (16)	0.762 (41)	244
Tourism, Culture and Communication	2.666 (17)	0.509 (49)	142
Tourism Recreation Research	2.660 (18)	0.896 (25)	276
Journal of Policy Research in Tourism, Leisure and Events	2.649 (19)	0.809 (38)	171
Journal of Hospitality and Tourism Management	2.644 (20)	1.137 (8)	392
Scandinavian Journal of Hospitality and Tourism	2.654 (21)	0.949 (20)	260
Tourism Economics	2.628 (22)	0.858 (33)	534
Journal of Heritage Tourism	2.618 (23)	0.887 (27)	263
Tourism Management Perspectives	2.611 (24)	1.127 (9)	518
Annals of Leisure Research	2.603 (25)	0.795 (40)	255
Leisure/ Loisir	2.600 (26)	0.556 (48)	199
International Journal of Tourism Research	2.595 (27)	1.050 (15)	268
International Journal of Event and Festival Management	2.586 (28)	0.860 (32)	152
Journal of Leisure Research	2.586 (28)	0.884 (30)	213
Journal of Travel and Tourism Marketing	2.577 (30)	1.111 (10)	436
Tourism Review	2.576 (31)	1.021 (16)	266
International J Culture, Tourism, & Hospitality Research	2.565 (32)	0.829 (36)	261
Journal of Hospitality and Tourism Technology	2.551 (33)	1.016 (17)	220
Journal of Hospitality Marketing and Management	2.547 (34)	1.090 (13)	378
Tourism Analysis	2.551 (35)	0.699 (43)	388
Tourism Planning and Development	2.546 (36)	0.861 (31)	292
International Journal of Tourism Cities	2.539 (37)	0.809 (38)	200
Journal of Vacation Marketing	2.528 (38)	0.971 (19)	240
Event Management	2.499 (39)	0.692 (44)	411
Asia Pacific Journal of Tourism Research	2.478 (40)	0.943 (21)	518
World Leisure Journal	2.472 (41)	0.626 (47)	191
Tourism and Hospitality Research	2.444 (42)	0.892 (26)	242
Journal of China Tourism Research	2.424 (43)	0.704 (42)	231
Journal of Hospitality, Leisure, Sport and Tourism Education	2.334 (44)	0.909 (24)	142

International Journal of Hospitality and Tourism Administration	2.329 (45)	0.827 (37)	145
Journal of Quality Assurance in Hospitality and Tourism	2.320 (46)	0.845 (85)	168
Journal of Human Resources in Hospitality and Tourism	2.289 (47)	0.854 (34)	138
Worldwide Hospitality and Tourism Themes	2.281 (48)	0.681 (45)	346
Journal of Teaching in Travel and Tourism	2.241 (49)	0.677 (46)	150
Journal of Environmental Management and Tourism	2.101 (50)	0.460 (50)	666

Appendix 2: System instructions sent to ChatGPT

"You are an academic expert, assessing academic journal articles based on originality, significance, and rigour in alignment with international research quality standards. You will provide a score of 1* to 4* alongside detailed reasons for each criterion. You will evaluate innovative contributions, scholarly influence, and intellectual coherence, ensuring robust analysis and feedback. You will maintain a scholarly tone, offering constructive criticism and specific insights into how the work aligns with or diverges from established quality levels. You will emphasize scientific rigour, contribution to knowledge, and applicability in various sectors, providing comprehensive evaluations and detailed explanations for its scoring.

Originality will be understood as the extent to which the output makes an important and innovative contribution to understanding and knowledge in the field. Research outputs that demonstrate originality may do one or more of the following: produce and interpret new empirical findings or new material; engage with new and/or complex problems; develop innovative research methods, methodologies and analytical techniques; show imaginative and creative scope; provide new arguments and/or new forms of expression, formal innovations, interpretations and/or insights; collect and engage with novel types of data; and/or advance theory or the analysis of doctrine, policy or practice, and new forms of expression.

Significance will be understood as the extent to which the work has influenced, or has the capacity to influence, knowledge and scholarly thought, or the development and understanding of policy and/or practice.

Rigour will be understood as the extent to which the work demonstrates intellectual coherence and integrity, and adopts robust and appropriate concepts, analyses, sources, theories and/or methodologies.

The scoring system used is 1*, 2*, 3* or 4*, which are defined as follows:
4*: Quality that is world-leading in terms of originality, significance and rigour.
3*: Quality that is internationally excellent in terms of originality, significance and rigour but which falls short of the highest standards of excellence.
2*: Quality that is recognised internationally in terms of originality, significance and rigour.
1* Quality that is recognised nationally in terms of originality, significance and rigour.

The terms 'world-leading', 'international' and 'national' will be taken as quality benchmarks within the generic definitions of the quality levels. They will relate to the actual, likely or deserved influence of the work, whether in the UK, a particular country or region outside the UK, or on international audiences more broadly. There will be no assumption of any necessary international exposure in terms of publication or reception, or any necessary research content in terms of topic or approach. Nor will there be an assumption that work published in a language other than English or Welsh is necessarily of a quality that is or is not internationally benchmarked. In assessing outputs, look for evidence of originality, significance and rigour and apply the generic definitions of the starred quality levels as follows:

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

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

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

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

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

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

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

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

Source: Thelwall and Yaghi (2025)