



OPEN A large language model pipeline for automated citation quality scoring across engineering journal quartiles with expert validation

Murat Isik & Sultan Guleryuzlu

This study proposes and evaluates a two-stage large language model (LLM)-based pipeline for automated citation quality scoring in academic manuscripts. The pipeline operates as follows: in Stage 1, citation sentences are extracted from full-text PDFs and matched to their referenced articles using the Gemini 2.5 Flash model; in Stage 2, each citation–reference pair is scored for semantic relevance on a continuous 0–10 scale by a second LLM inference call operating under a structured five-tier rubric and a skeptical reviewer prompt persona. The pipeline was applied to a corpus of 121 Web of Science (WOS)-indexed engineering articles drawn from journals spanning all four Journal Citation Reports quartile strata (Q1–Q4), yielding 5,615 scored citation–reference pairs. Descriptive analysis revealed an overall mean relevance score of 7.76 (SD = 2.36), with 74.7% of citations rated as Strong or Excellent. A Kruskal–Wallis test confirmed statistically significant score differences across quartile groups ($H(3) = 157.10$, $p < 0.001$), though the overall effect size was small ($\epsilon^2 = 0.028$). Post-hoc Mann–Whitney U tests with Bonferroni correction identified Q2 articles as recording the highest mean scores ($M = 8.04$), significantly outperforming Q1 ($M = 7.52$), Q3 ($M = 7.73$), and Q4 ($M = 7.74$). The Q3 versus Q4 comparison was the sole non-significant pairing ($p = 0.756$), indicating these strata are statistically indistinguishable in citation quality. Spearman correlation yielded a weak negative rank correlation ($\rho = -0.105$, $p < 0.001$), with Q1 recording the highest proportion of Irrelevant citations (10.7%). These findings challenge the assumption that citation quality improves monotonically with journal prestige. The lower mean score of Q1 coexists with one of the highest proportions of highly relevant citations, indicating a bimodal rather than uniformly weaker profile, and a systematic annotation showed that context-dependent pointer citations are disproportionately concentrated in the Q1 Irrelevant set. We therefore attribute Q1's pattern to the broader interdisciplinary scope of top-tier articles together with a measurement effect, rather than to any single cause such as AI-assisted writing. The proposed pipeline offers a scalable, content-aware complement to existing academic integrity tools, with practical applications in editorial pre-screening and automated peer review support. An inter-rater reliability study on a stratified subsample of 150 citation–reference pairs showed strong ordinal agreement between the LLM and expert majority vote (Spearman $\rho = 0.643$, $p < 0.001$), with exact-category agreement of 48.0% rising to 77.3% under ± 1 adjacent-category tolerance, and highest agreement at the Irrelevant (80.0%) and Excellent (71.0%) poles.

Keywords Citation analysis, Scientific publishing, Natural language processing, Text analysis, Machine learning

Citations constitute the foundational infrastructure of scholarly communication, enabling researchers to situate new findings within existing knowledge, acknowledge intellectual debts, and establish the evidential basis for scientific claims. The volume of academic literature has grown at an unprecedented pace: according to¹, approximately 2.82 million journal articles were indexed in major databases in 2022 alone, representing a 47% increase compared to 2016. This expansion has placed severe strain on the peer review system, eroding the time and cognitive resources that reviewers can devote to any individual manuscript^{1,2}. One consequence that has

Department of Computer Engineering, Kirsehir Ahi Evran Universitesi, 40100 Kirsehir, Turkey. ✉email: muratisik@ahievran.edu.tr

received comparatively little systematic attention is the quality of individual citations — specifically, whether the sources cited in a manuscript substantively support the claims they are invoked to justify.

The widespread adoption of LLMs in academic writing adds a qualitatively new dimension to this challenge. LLMs generate text by interpolating statistical patterns in training data rather than by retrieving or consulting actual documents or papers. As a consequence, they routinely produce citations that are factually non-existent, bibliographically malformed, or semantically unrelated to the claim they ostensibly support^{3,4} — a phenomenon that falls within the broader category of hallucination in natural language generation⁵. High-profile cases have made this risk visible beyond the research community: in one widely reported incident, attorneys submitted AI-generated court briefs that cited non-existent legal precedents⁶. In the scientific domain, empirical studies have documented hallucination rates of approximately 20% overall, with rates ranging from about 3% to over 45% depending on topic familiarity, prompt specificity, and subject domain^{7,8}. At scale, a benchmark analysis examined 375,440 citations. These citations were generated by 13 state-of-the-art LLMs. The analysis found hallucination rates ranging from 14.23% to 94.93% across models. A separate archival study examined over 56,000 peer-reviewed papers. That study confirmed an 80.9% increase in publications containing invalid citations in 2025⁹. Recent investigations have documented hallucinated citations in peer-reviewed publications at prestigious AI venues: one report identified more than 100 AI-hallucinated references across 53 accepted papers at NeurIPS 2025¹⁰, while a separate investigation found that 50 papers submitted to ICLR 2026 each contained at least one hallucinated citation that had passed peer review¹¹.

Current approaches to detecting AI-generated academic content rely predominantly on stylometric and linguistic features — statistical properties of text such as perplexity and token predictability — as implemented in tools such as GPTZero, Turnitin AI, and ZeroGPT, whose detection performance has been systematically evaluated¹². These tools offer scalable screening but are known to produce false positives at non-trivial rates, particularly for non-native English speakers, and remain vulnerable to evasion through paraphrasing attacks^{13,14}. Critically, stylometric detectors operate on the surface properties of text and cannot assess the semantic relationship between a citation sentence and the document it references. Citation quality scoring offers a structurally different and complementary mode of verification: rather than analysing how a text is written, it probes whether what is claimed is actually supported by the cited source.

Prior work on citation analysis has established that citations are not equally important or evidentially meaningful. Pioneering work by¹⁵ introduced the concept of citation strength estimation and demonstrated that not all citations carry equal importance, a finding corroborated by subsequent research distinguishing important from incidental citations¹⁶. More recent work has extended citation context analysis through deep learning and pre-trained language model methods¹⁷, and automated verification tools such as SemanticCite¹⁸ and SourceCheckup¹⁹ have demonstrated that LLM-based pipelines can assess citation support at scale. However, no prior study has applied continuous LLM-based quality scoring to a stratified sample of peer-reviewed engineering articles to examine whether citation quality varies systematically across journal prestige strata as indexed by Web of Science (WOS) quartile rankings. This represents the central empirical gap that the present study addresses.

The present study proposes and evaluates a two-stage automated pipeline for citation quality scoring. In Stage 1, citation sentences are extracted from full-text PDFs using the Gemini 2.5 Flash model and matched to their corresponding referenced articles. In Stage 2, each citation–reference pair is scored for semantic relevance on a continuous 0–10 scale by a second Gemini 2.5 Flash inference call operating under a structured rubric and a “skeptical reviewer” prompt persona, consistent with best practices in LLM-as-a-judge methodology^{20,21}. The pipeline was applied to a corpus of 121 WOS-indexed engineering articles drawn from journals spanning all four quartile strata (Q1–Q4), yielding 5,615 scored citation–reference pairs. The study addresses three specific research questions: (1) Is LLM-based citation quality scoring methodologically valid as a large-scale assessment instrument? (2) Do citation quality scores differ significantly across WOS quartile strata? (3) Is there a monotonic relationship between journal prestige rank and citation quality?

This study makes three principal contributions to the literature. First, it introduces and operationalises a continuous citation quality scoring framework deployable at corpus scale, extending prior binary or categorical verification approaches^{18,19}, and provides initial evidence of agreement with expert human judgment through an inter-rater reliability study comparing LLM classifications with independent expert evaluations. Second, it provides the first large-scale empirical evidence on the relationship between WOS journal quartile and citation quality in the engineering domain, revealing a non-monotonic pattern in which Q2 journals record higher mean citation quality scores than Q1 journals. Third, it demonstrates that citation quality scoring can function as an indirect, complementary signal for detecting anomalous bibliographic practices — including those associated with LLM-assisted writing — within a multi-modal academic integrity framework. Given the accelerating integration of LLMs into scholarly workflows^{13,22} and the documented strain on the peer review system that limits reviewers’ capacity to verify individual citations^{1,2}, the development of scalable, automated citation quality assessment tools represents a practical need as well as a methodological advance.

The remainder of this paper is organised as follows. The section “[Literature review](#)” reviews the relevant literature across four areas: citation quality and semantic relevance, LLM-based citation annotation and verification, hallucinated citations in the era of LLM-assisted writing, and the relationship between journal prestige and citation practices. The section “[Methodology](#)” describes the methodology, including corpus construction, the two-stage pipeline architecture, the scoring rubric, the inter-rater reliability design, and the statistical analysis. The section “[Results](#)” presents the results of descriptive and inferential statistical analyses, together with the inter-rater reliability and sensitivity analyses. The section “[Discussion](#)” discusses the findings in relation to the existing literature. The section “[Limitations](#)” outlines the limitations, the section “[Implications and future directions](#)” presents implications and future directions, and the section “[Conclusion](#)” concludes.

Literature review

Citations are the connective tissue of scientific discourse, linking new findings to prior knowledge and establishing the evidential basis for scholarly claims. Yet the mere presence of a citation does not guarantee that the referenced source substantively supports the assertion it accompanies. This observation has motivated a long line of inquiry into citation quality — broadly understood as the degree to which a citation sentence accurately and relevantly invokes its referenced material. Early bibliometric approaches relied on citation counts as a proxy for scientific impact, but such aggregate metrics are blind to the semantic relationship between citing and cited texts — a limitation that has motivated a growing body of NLP-based in-text citation analysis research over the past decade²³. Sustained criticism of this limitation gave rise to citation context analysis, a subfield dedicated to examining the textual environment in which individual citations appear²³, with recent work advancing segment-level and LLM-based approaches to the task^{24,25}.

Within citation context analysis, two analytical traditions have developed in parallel. The first focuses on citation function classification — categorising citations according to the role they play in the argument structure of the citing paper (e.g., background, method, comparison, or supportive)^{17,24}. Influential annotation schemes such as those of^{26,27} have been widely adopted and refined, and deep learning models have progressively improved classification performance on standard benchmarks¹⁷. The second tradition focuses on citation relevance or importance — asking not how a source is used but how well the cited material supports the claim being made. It is this second tradition that the present study most directly extends.

Pioneering work by¹⁵ introduced the concept of citation strength estimation, demonstrating that individual citations differ substantially in the importance they carry, a finding corroborated by subsequent research on important versus incidental citations¹⁶. More recently, the growing availability of large pre-trained language models has expanded the methodological toolkit available for semantic citation analysis. Zhang¹⁷, provide a comprehensive survey of the transition from feature-engineering approaches to deep learning and large language model (LLM)-based methods in citation function classification, documenting both the rapid gains in accuracy and persistent challenges related to annotation scarcity and class imbalance. The present study draws on this methodological progression by deploying a frontier LLM — Gemini 2.5 Flash — as a semantic scoring agent for citation–reference pairs, rather than a citation function classifier.

The use of LLMs as annotation tools in citation context analysis is an emerging but rapidly growing area. Nishikawa and Koshiba²⁵, provide the most directly comparable prior study: they investigated the applicability of GPT-3.5-turbo to the task of classifying citation purpose and sentiment by comparing LLM outputs to expert human annotations. Their results revealed an important asymmetry — LLM annotations were comparable to or better than human annotations in consistency, yet performed poorly in terms of predictive performance relative to gold-standard classifications. The authors concluded that immediate replacement of human annotators by LLMs was inappropriate, but that LLM outputs could serve as useful reference information to support and accelerate human annotation workflows. This finding has direct implications for the present study: rather than framing the LLM as a replacement for expert judgment, our pipeline positions it as a structured scoring instrument operating within a carefully designed rubric, a design choice intended to harness LLM consistency while mitigating the predictive accuracy shortfall documented by Nishikawa and Koshiba.

Haan¹⁸, has introduced a complementary perspective named SemanticCite, an AI-powered framework for automated full-text citation verification. Unlike approaches that rely on abstract-level matching or binary exist/non-exist classification, SemanticCite employs a four-stage pipeline as follows; text preprocessing, hybrid retrieval, neural reranking, and LLM-based analysis. The task is to assign each citation to one of four support categories: Supported, Partially Supported, Unsupported, and Uncertain. Applied to a dataset of 1,111 citations across eight academic disciplines, the framework demonstrated that fine-tuned lightweight language models can achieve classification performance comparable to large commercial systems at substantially lower computational cost. SemanticCite shares the present study's core premise that the semantic relationship between a citation sentence and its referenced document can be assessed automatically, but differs in scope and objective: it is designed as a general-purpose verification tool rather than a comparative study of citation quality across journal prestige strata.

The broader applicability of LLMs as judges of semantic quality is supported by²¹, who demonstrated that strong LLM judges such as GPT-4 can achieve agreement with human evaluators at rates exceeding 80%, a level comparable to inter-human agreement, when evaluated on general-purpose language tasks. Thakur et al.²⁰, extended this line of work by systematically documenting the biases to which LLM judges are susceptible — including verbosity bias and position bias — and showing that structured prompting strategies can substantially mitigate these effects. Taken together, these studies underpin the methodological validity of the LLM-as-a-judge paradigm that our pipeline instantiates, while also motivating the specific prompt design choices described in the section “[Methodology](#)”.

A parallel and rapidly expanding literature documents the risks that LLM-generated text poses to the integrity of academic citations. Because LLMs generate plausible-sounding references by interpolating patterns in training data rather than retrieving records from indexed databases, they frequently produce citations that are factually non-existent, bibliographically malformed, or semantically misaligned with the surrounding text — a phenomenon variously termed “hallucinated citations,” “ghost citations,” or “fabricated references”^{3–5}.

Early case-report evidence of this problem surfaced prominently in legal and clinical contexts. Neumeister⁶, reported the widely publicised case in which attorneys submitted court briefs citing non-existent legal precedents generated by ChatGPT, while³ demonstrated that ChatGPT could generate plausible but entirely fabricated scientific references when asked to provide sources on specific medical topics such as homocysteine-induced osteoporosis. These observations were quickly followed by systematic empirical investigations. Mugaanyi et al.⁸, prompted ChatGPT 3.5 to write introduction sections across ten manuscript topics spanning the natural sciences and humanities, generating a total of 102 citations. Of these, 76 (74.5%) were verified to exist in the

published literature; existence rates did not differ significantly between natural sciences (72.7%) and humanities (76.6%), but accuracy rates — defined as exact bibliographic correctness — were substantially higher in the natural sciences (32.7%) than in the humanities (8.5%), indicating that citation existence and citation accuracy are distinct dimensions of quality that can diverge markedly across disciplinary contexts.

The scale of the hallucination problem has been further quantified by⁹, who developed CiteVerifier, an open-source framework for large-scale citation validity assessment, and applied it in three complementary experiments. A benchmark of 13 state-of-the-art LLMs across 40 research domains, covering 375,440 generated citations, found that all models hallucinated citations at rates ranging from 14.23% (DeepSeek) to 94.93% (Hunyuan), with substantial variation across domains. An archival analysis of 2.2 million citations from 56,381 papers published at top-tier AI/ML and security venues between 2020 and 2025 confirmed that 1.07% of papers contained invalid or fabricated citations, with an 80.9% increase in affected papers in 2025 alone. A researcher survey further revealed a critical verification gap: 41.5% of respondents reported copying BibTeX entries without verification, and 76.7% of reviewers acknowledged not thoroughly checking references. These findings collectively establish the magnitude and trajectory of a problem that the present study's scoring pipeline is designed, in part, to address.

It is important to distinguish the hallucination problem studied by^{8,9} — which concerns citations generated by LLMs during text production — from the citation quality problem studied in the present work, which concerns citations written by human authors. Our dataset consists entirely of human-authored engineering articles indexed in WOS, and the LLM is used not to generate citations but to evaluate them. The two phenomena are, however, causally linked: as LLM assistance in academic writing proliferates^{13,22}, verification tools capable of assessing the semantic appropriateness of citations in human-authored manuscripts become increasingly important, regardless of whether those citations were LLM-suggested or independently selected by the author.

The recognition that citation inaccuracy represents a systemic threat to scholarly communication has spurred the development of automated verification frameworks. Beyond SemanticCite¹⁸, described above, a notable contribution is SourceCheckup¹⁹, an automated framework developed specifically for the medical literature. SourceCheckup evaluates the degree to which LLM-generated answers are supported by the cited medical references by decomposing responses into individual citation-sentence pairs and assessing evidential support automatically. Applied to approximately 800 clinical questions generating over 58,000 sentence-source pairs, the system revealed that citation support in LLM medical outputs varies considerably and that a non-trivial proportion of cited claims are not adequately grounded in the referenced sources. The SourceCheckup framework is architecturally analogous to Stage 2 of our pipeline — both map citation sentences to referenced documents and score support — but targets LLM output rather than human-authored manuscripts and operates in a domain-specific clinical context rather than across engineering disciplines.

The methodological space occupied by these verification tools can be located within the broader landscape of automated scholarly paper review (ASPR), surveyed comprehensively by²⁸. Drawing on a snowball sample of the ASPR literature²⁸, document the rapid proliferation of LLM-based systems for tasks ranging from manuscript screening and compliance checking to full review generation. Their survey identifies citation verification as a component of emerging ASPR architectures but notes that most existing systems treat citation checking as a secondary function rather than as a primary analytical objective. The present study addresses this gap by placing citation quality assessment at the centre of the pipeline and operationalising it as a quantitative scoring problem amenable to comparative statistical analysis.

The relationship between journal prestige and the quality of scholarly output — including citation practices — has been examined from multiple angles. The standard assumption, embedded in the hierarchical logic of journal impact factors and quartile rankings, is that higher-tier outlets impose more rigorous review standards, attracting manuscripts that are better executed across all quality dimensions including the accuracy and relevance of citations. Hanson et al.¹, document the increasing strain on the peer review system arising from publication volume growth, reviewer shortages, and shortened review timescales, raising the possibility that even high-prestige journals face mounting pressure on quality control. Donker²², elaborated on the specific risks posed by the use of LLMs in the peer review process, arguing that LLM-assisted reviews may introduce systematic blind spots precisely in the domain of citation verification.

Whether these systemic pressures manifest differentially across journal prestige strata is an empirical question that has received limited direct investigation. Studies comparing citation accuracy or relevance across WOS quartiles are notably absent from the literature, leaving open the question of whether the quality advantage of higher-ranked journals extends to the semantic appropriateness of individual citations. The present study provides the first large-scale quantitative evidence bearing on this question in the engineering domain, finding that the relationship between journal prestige and citation quality is non-monotonic and more nuanced than the standard assumption would suggest.

The review above reveals three converging gaps in the existing literature. First, while LLM-based tools have been applied to citation function classification²⁵ and large-scale citation validity assessment⁹, and empirical studies have documented the extent of citation fabrication in LLM-generated text⁸, no prior study has employed an LLM as a semantic quality scorer for citation-reference pairs drawn from a stratified sample of peer-reviewed engineering articles. Second, automated citation verification tools such as SemanticCite¹⁸ and SourceCheckup¹⁹ operate as binary or categorical classifiers and do not produce continuous quality scores that permit parametric or non-parametric comparison across groups. Third, the relationship between journal prestige and citation quality — a question with direct implications for the interpretation of bibliometric indicators — has not been subjected to systematic empirical investigation. The present study addresses all three gaps simultaneously: it introduces a two-stage LLM pipeline that assigns a continuous quality score to each citation-reference pair, applies this pipeline to 5,615 pairs drawn from 121 WOS-indexed engineering articles stratified by quartile, and uses non-parametric inferential statistics to test whether citation quality varies across prestige strata.

Methodology

This study proposes and evaluates an LLM-based automated pipeline for assessing the semantic relevance between citation sentences and their corresponding reference articles. The framework addresses a critical gap in citation quality research: the absence of scalable, content-aware tools capable of detecting superficial or potentially hallucinated citations in academic manuscripts. The pipeline consists of two sequential stages — citation sentence extraction and relevance scoring — both implemented using large language model inference via API calls.

Corpus construction

Journal selection

The sampling procedure was initiated using the Web of Science (WOS) Master Journal List, from which only journals indexed under Engineering disciplines were retained. Engineering was selected as the single target field for several reasons. First, citation norms, reference density, and the typical rhetorical function of citations vary substantially across broad disciplinary domains; restricting the corpus to one field controls for this between-field heterogeneity, so that any observed differences in citation quality can be attributed to the journal prestige stratum rather than to disciplinary citation culture. Second, engineering articles typically follow a standardized structure with explicit, numbered reference lists, which is well suited to the automated citation-sentence extraction and citation–reference matching performed in Stage 1. Third, the validation of the pipeline depended on volunteer expert evaluators competent to judge citation relevance, all of whom were faculty members in engineering and related computing disciplines; confining the corpus to engineering therefore ensured that the citation–reference pairs fell squarely within the evaluators’ domain expertise, which was essential for a meaningful inter-rater reliability study. Engineering is also among the largest and most finely quartile-stratified categories in the Web of Science, providing an ample article pool across all four strata. Extending the analysis to other disciplines is an important direction for future work. A lightweight automated tool was developed to enrich this list by appending the Journal Citation Reports (JCR) quartile designation (Q1–Q4) to each journal entry. From the resulting quartile-stratified journal pool, journals were selected through simple random sampling, with an approximately equal number of journals drawn from each quartile stratum to ensure balanced representation.

Article selection

From each randomly selected journal, up to two eligible articles published between 2020 and 2024 were retrieved, subject to availability under the inclusion criteria. To maintain corpus homogeneity and minimize confounding factors, the following inclusion criteria were applied:

- The article must be an original research article (reviews, editorials, letters, and corrections were excluded).
- The article must be written in English.
- The article must be available in full-text PDF format.

This procedure yielded a final analyzable corpus of 121 articles across the four quartile strata (Q1: 29, Q2: 31, Q3: 30, Q4: 31).

Reference filtering

Each article in the corpus contained a reference list. The analysis used only references to peer-reviewed journal articles. It excluded non-journal sources. These sources included conference proceedings, books, book chapters, theses, technical reports, and web pages. In this way, it was ensured that each citation–reference pair consisted of two comparable, full-text scholarly documents suitable for semantic evaluation.

Automated pipeline

Stage 1: citation sentence extraction

For each source article, citation sentences were extracted automatically using the Gemini 2.5 Flash model accessed via the Google Generative AI API. The input to the model consisted of the full text of the article, extracted from its PDF using the PDF.js library (client-side extraction), along with the list of reference titles provided by the user. The extraction prompt was engineered with the following non-negotiable directives:

- **Completeness:** Every citation in the main body text must be identified, with no omissions.
- **Figure/table exclusion:** Citations appearing exclusively within figure captions, table titles, or image descriptions were discarded.
- **PDF artifact correction:** Common PDF-to-text conversion errors (e.g., broken hyphenation such as “classi fi cation”) were identified and corrected prior to extraction.
- **Handling of multiple citations:** For sentences containing multiple co-citations or citation ranges (for example, several consecutive references cited together), the original citation sentence was retained verbatim and paired separately with each individual cited reference, yielding one citation–reference pair per cited source. Each pair was then scored independently in Stage 2, so that the same sentence could receive different relevance scores with respect to different references. Citation ranges were expanded only at the level of reference enumeration; the wording of the citation sentence was never altered or regenerated by the model.
- **Reference matching:** Each extracted citation was linked to its corresponding source article title through a three-step reasoning process: (i) identifying the bibliography entry associated with the citation marker, (ii) semantically matching the bibliography title to the user-supplied reference list, and (iii) verbatim copying of the matched title from the user list to preserve original formatting.

The structured output was enforced via JSON Schema, producing for each citation a citation_sentence field and a source_name field.

Stage 2: relevance scoring

Each citation–reference pair was subsequently scored for semantic relevance using a second Gemini 2.5 Flash inference call. The input consisted of the citation sentence and the full text of the corresponding reference article (extracted from PDF). The model was assigned the persona of a rigorous academic reviewer with an explicitly skeptical prior disposition, and was instructed to score the relevance on a continuous scale from 0.0 to 10.0 according to a five-tier rubric as shown in Table 1.

The rubric operationalizes citation relevance as the degree of semantic centrality between the citation sentence and the content of the cited article. To put it another way, how closely the citation engages the cited work’s substantive content, graded from its core opinion (Excellent) through a significant or extensively elaborated element (Strong) to a peripheral or background detail (Moderate). This gradient reflects centrality of engagement, not the appropriateness or correctness of the citation per se: a citation to a secondary finding, a methodological protocol, or a reused dataset can be entirely appropriate, and is captured at the Strong level when it substantively engages content elaborated in the cited article, or at the Moderate level when it refers to a more peripheral element. Only the Irrelevant band (0.0–2.9) denotes a genuinely problematic citation, in which no meaningful semantic connection exists or the reference contradicts the citing claim. Intermediate scores should therefore be interpreted as indicating lower centrality of engagement rather than inferior citation practice.

The category boundaries were placed at uniform two-point intervals on the underlying continuous scale, at cut points of 2.9, 4.9, 6.9, and 8.9. The resulting differences in band width arise only from the bounds of the 0–10 scale: the Excellent band is narrow (8.9–10.0) because it is capped by the scale ceiling, whereas the Irrelevant band is wide (0.0–2.9) because it extends to the scale floor. These choices are also deliberate. The narrow Excellent band raises the threshold for the top category, reinforcing the anti-inflation design by reserving the highest rating for citations that engage the cited work’s core contribution, while the wide Irrelevant band consolidates all citations lacking a meaningful semantic connection into a single, actionable category suited to editorial flagging. Crucially, all inferential analyses were conducted on the continuous 0–10 scores rather than on the categorical bands, so the band widths affect only the interpretive presentation of the results and not the statistical tests or effect sizes.

To counteract the well-documented score inflation tendency of generative language models^{20,21}, the prompt explicitly instructed the model to maintain a broad score distribution and to treat each citation independently. Scores were clipped to the [0, 10] interval as a post-processing step. The structured JSON output enforced two fields: score (numeric) and justification (a concise, evidence-based rationale in Turkish).

A delay of 1,500 milliseconds was introduced between consecutive API calls to comply with rate limiting constraints.

Dataset characteristics

Model configuration and reproducibility. Both pipeline stages used Google’s gemini-2.5-flash model accessed through the @google/genai SDK, with structured JSON output enforced by a response schema (Stage 1: an array of {citation_sentence, source_name}; Stage 2: an object of {score, justification}). Generation parameters (temperature, top_p, top_k, and max_output_tokens) were left at the API’s default values and no random seed was fixed. The Stage 1 extraction-and-matching prompt was written in English and the Stage 2 scoring prompt in Turkish (the working language of the author team), with the model instructed to return Turkish justifications. Both prompt templates are provided verbatim in the public code repository (see Code Availability). Representative scored examples for all five rubric categories across the four quartile strata are shown in Table 2.

The complete pipeline processed a total of 8,382 citation–reference pairs. Of these, 5,615 pairs (67.0%) were successfully scored. A further 1,752 pairs (20.9%) were excluded at the reference filtering stage, as they corresponded to non-journal sources — including conference proceedings, books, book chapters, theses, and technical reports — which were ineligible for scoring under the criteria described above. The remaining 1,015 pairs (12.1%) could not be processed due to inaccessibility of the reference article in full-text PDF format, attributable to the absence of open-access availability and the limitations of the institutional subscription agreements. The final analyzable dataset thus comprised 5,615 citation–reference pairs drawn from 121 source articles across four JCR quartile strata.

Score Range	Category	Criterion
9.0–10.0	Excellent	The citation sentence reflects the primary thesis, principal finding, or central argument of the reference article
7.0–8.9	Strong	The citation accurately represents a significant finding, secondary result, or a concept extensively elaborated in the reference
5.0–6.9	Moderate	The citation relates to a peripheral topic, background information, or a less prominent detail of the reference
3.0–4.9	Weak	The connection is indirect or interpretive; the citation shares only a keyword with the reference without capturing its core message
0.0–2.9	Irrelevant	No meaningful connection exists, or the reference contradicts the citing claim

Table 1. Five-Tier Rubric Metric.

Quartile	Category	Score	Citation_Sentence	Reference Content Summary
Q1	Irrelevant	1	The work by [ref] cannot explicitly define the target model size.	A filter-level pruning method that compresses deep neural networks by removing redundant convolutional channels.
Q2	Weak	4.1	In recent years the application of deep learning in computer vision has been steadily rising due to its versatility and performance [ref] .	A deep architecture that learns hierarchical features directly from unordered 3D point-cloud data.
Q3	Moderate	6.5	However, compared to 2D convolution [ref] , a 3D convolution of sequences of dense frames is computationally expensive, and the model size also has a quadratic growth.	An attention module that sequentially applies channel and spatial attention within convolutional networks.
Q4	Strong	8.3	A lot of work has addressed the application of GAs especially for medical image registration [ref].	A genetic-algorithm approach for robust three-dimensional MR-to-CT medical image registration.
Q1	Excellent	9.8	Tsai et al. proposed to use local geometric characteristics of vertices to achieve integrity protection [ref].	A low-complexity, region-based authentication algorithm that protects the integrity of 3D polygonal mesh models.

Table 2. Representative Scored Citation Examples, One Per Rubric Category, with Neutral Reference Content Summaries.

Inter-rater reliability study design

To benchmark the LLM-based scoring pipeline against expert human judgment, an inter-rater reliability (IRR) study was conducted. A stratified random subsample of 150 citation–reference pairs was drawn from the scored dataset, with 30 pairs selected from each of the five rubric categories (Irrelevant, Weak, Moderate, Strong, Excellent) to ensure adequate representation of all quality levels, including rare categories. Three independent evaluators, each holding a faculty position — two from the Department of Computer Engineering at Kirsehir Ahi Evran University and one from the Department of Management Information Systems at Sakarya University, independently classified each pair into one of the five rubric categories defined in Table 1. Evaluators were provided with the citation sentence and the referenced article title, and were blind to the LLM-assigned score and category. The order of pairs was randomised to prevent evaluators from inferring the underlying category distribution.

Inter-expert agreement was quantified using Fleiss' κ for multi-rater nominal agreement and Cohen's κ for pairwise comparisons. Expert–LLM agreement was assessed both at the individual expert level and using the expert majority vote as the consensus label. Spearman's rank-order correlation was computed for all pairwise comparisons to capture ordinal agreement. Per-class exact-match and ± 1 adjacent-category agreement rates were calculated to evaluate whether disagreements reflected gross misclassifications or boundary effects between adjacent categories. A confusion matrix was constructed to visualise the distribution of agreements and disagreements across the five quality categories.

Informed consent was obtained from all three expert evaluators prior to their participation in the inter-rater reliability study. Evaluators were informed about the purpose of the study, the nature of the evaluation task, and the use of their assessments in the published manuscript.

Statistical analysis

Given that the Shapiro–Wilk test confirmed non-normal score distributions across all four quartile groups (all $p < 0.001$), non-parametric methods were employed throughout. Group differences were assessed using the Kruskal–Wallis H test, followed by pairwise Mann–Whitney U tests with Bonferroni correction (adjusted $\alpha = 0.0083$). Effect sizes were quantified using the rank-biserial correlation coefficient (r) for pairwise comparisons and epsilon-squared (ϵ^2) for the omnibus Kruskal–Wallis test. The monotonic relationship between journal quartile rank and citation relevance score was evaluated using Spearman's rank-order correlation. 95% confidence intervals for all effect sizes (ϵ^2 , rank-biserial r , and Spearman's ρ) were estimated by stratified bootstrap resampling (5,000 resamples, percentile method). All statistical analyses were conducted in Python 3 using the SciPy library.

Results

Dataset overview

The final analyzable dataset comprised 5,615 citation–reference pairs drawn from 121 source articles distributed across four WOS quartile strata (Q1: 29 articles, $n = 1,838$ pairs; Q2: 31 articles, $n = 1,801$ pairs; Q3: 30 articles, $n = 1,077$ pairs; Q4: 31 articles, $n = 899$ pairs). A total of 1,015 citation–reference pairs (12.1% of the raw dataset) could not be scored due to inaccessibility of the reference article in full-text PDF format. Additionally, a small number of anomalies were identified in the Q4 stratum during data collection: one article contained a duplicate reference registered under two distinct citation numbers, and four citation instances across three articles pointed to non-existent sources. These non-existent-reference anomalies were retained as bibliographic anomalies and assigned a score of 0.0, rather than being treated as inaccessible full-text cases; they are discussed further in the section “Discussion”.

Overall score distribution

Descriptive statistics for the full dataset and each quartile stratum are presented in Table 3. Across all 5,615 scored citation–reference pairs, the mean relevance score was 7.761 (SD = 2.358) with a median of 8.5 (IQR = [6.8, 9.5]). The distribution was markedly left-skewed, with the majority of citations receiving high relevance scores. As shown in Fig. 1, the kernel density estimates for all four quartile groups exhibited a pronounced peak in the 8–10 score range, confirming the dominance of high-relevance citations across the corpus. Figure 2 further illustrates

Quartile	N	Mean	SD	Median	IQR	Min	Max
Q1	1,838	7.522	2.806	8.8	[6.5, 9.5]	0.0	10.0
Q2	1,801	8.039	2.238	8.9	[7.2, 9.6]	0.0	10.0
Q3	1,077	7.726	2.000	8.5	[7.0, 8.8]	0.0	10.0
Q4	899	7.737	1.882	8.2	[7.0, 9.0]	0.0	10.0
Overall	5,615	7.761	2.358	8.5	[6.8, 9.5]	0.0	10.0

Table 3. Descriptive Statistics of Citation Relevance Scores by WOS Quartile.

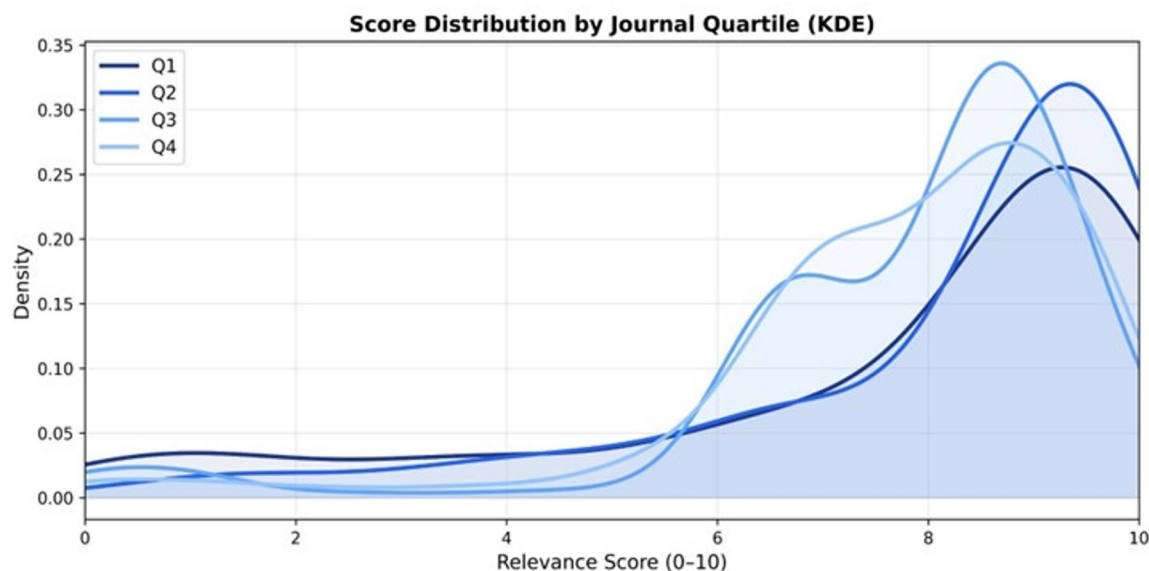


Fig. 1. Score Distribution by Journal Quartile (Kernel Density Estimation). Each curve represents the probability density of relevance scores for citations from articles in the corresponding WOS quartile.

these distributional differences through boxplots, highlighting the higher median and narrower interquartile range of Q2 relative to the other quartiles, as well as the greater prevalence of low-score outliers in Q1.

Citation quality category distribution

Table 4 presents the distribution of citation–reference pairs across the five quality categories defined by the scoring rubric. At the corpus level, 74.7% of all citations received scores in the Strong (35.2%) or Excellent (39.5%) range, indicating that the large majority of citations in WOS-indexed engineering articles are substantively relevant to their cited sources. Conversely, 11.6% of citations were rated as Irrelevant (6.7%) or Weak (4.9%), representing cases where the semantic connection between the citing sentence and the referenced article was limited or absent.

Notable differences in category distributions were observed across quartiles, as illustrated in Fig. 3. Q1 and Q2 articles showed the highest proportions of Excellent-rated citations (46.2% and 48.2%, respectively), whereas Q3 and Q4 articles were predominantly concentrated in the Strong category (51.2% and 49.4%). Conversely, Q1 exhibited the highest proportion of Irrelevant citations (10.7%), more than double the rate observed in Q3 (5.3%). Q3 and Q4 recorded the lowest overall rates of problematic citations (Irrelevant + Weak = 6.2% and 6.0%, respectively), while Q2 showed the lowest rate among the upper two quartiles (11.5%).

Inferential statistical analysis

Assumption testing

Prior to group comparisons, the normality of score distributions was assessed using the Shapiro-Wilk test on random subsamples of $n = 500$ per quartile group. All four groups significantly deviated from normality (Q1: $W = 0.788$, $p < 0.001$; Q2: $W = 0.774$, $p < 0.001$; Q3: $W = 0.716$, $p < 0.001$; Q4: $W = 0.807$, $p < 0.001$), as also evidenced by the bimodal and left-skewed distributions visible in Fig. 1. Accordingly, non-parametric statistical methods were employed for all subsequent comparisons.

Overall group differences

A Kruskal-Wallis H test revealed a statistically significant difference in citation relevance scores across the four quartile groups ($H(3) = 157.10$, $p < 0.001$, $\epsilon^2 = 0.028$, 95% CI [0.021, 0.037]). Although the overall effect size was small, the result confirms that journal quartile stratum is a statistically significant factor associated with citation relevance score variation.

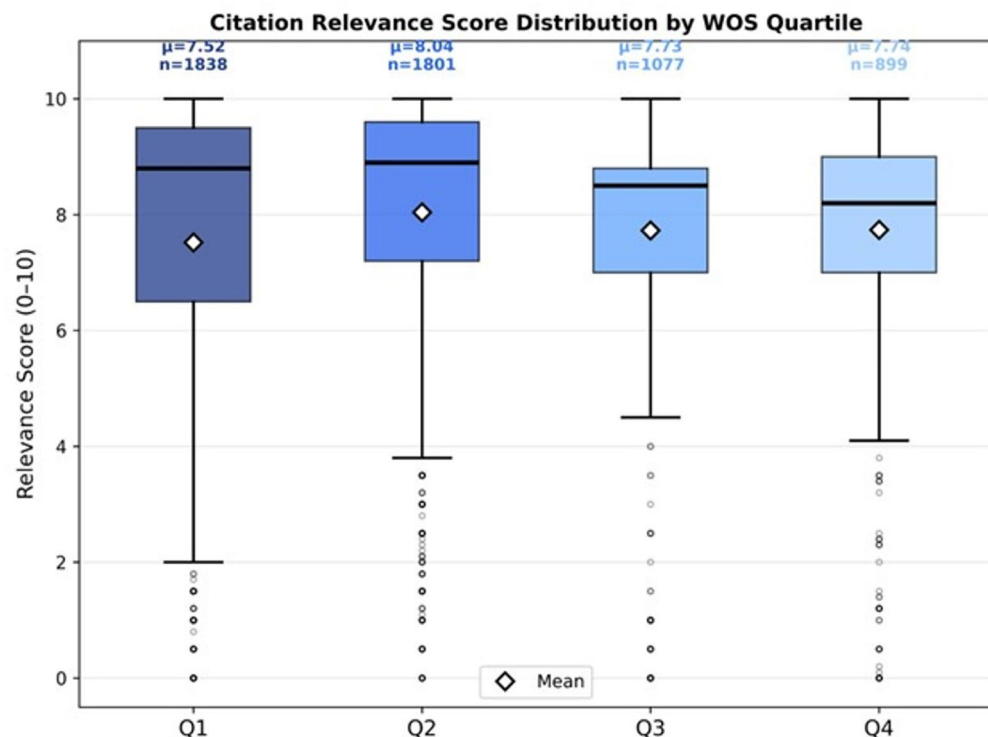


Fig. 2. Citation Relevance Score Distribution by WOS Quartile (Boxplot). Diamonds indicate group means; horizontal lines indicate medians; circles denote outliers.

Q	Irrelevant (0–2.9)	Weak (3–4.9)	Moderate (5–6.9)	Strong (7–8.9)	Excellent (9–10)	Total
Q1	196 (10.7%)	130 (7.1%)	196 (10.7%)	466 (25.4%)	850 (46.2%)	1,838
Q2	89 (4.9%)	118 (6.6%)	210 (11.7%)	516 (28.7%)	868 (48.2%)	1,801
Q3	57 (5.3%)	10 (0.9%)	199 (18.5%)	551 (51.2%)	260 (24.1%)	1,077
Q4	36 (4.0%)	18 (2.0%)	162 (18.0%)	444 (49.4%)	239 (26.6%)	899
Total	378 (6.7%)	276 (4.9%)	767 (13.7%)	1,977 (35.2%)	2,217 (39.5%)	5,615

Table 4. Citation Quality Category Distribution by WOS Quartile (Frequencies And Percentages).

Pairwise comparisons

Post-hoc pairwise Mann-Whitney U tests with Bonferroni correction (adjusted $\alpha=0.0083$) were conducted for all six quartile pairs. Results are presented in Table 5. Five of six comparisons yielded statistically significant differences. The Q2 stratum showed significantly higher scores than all other quartiles (Q2 vs. Q3: $r=0.248$, $p<0.001$; Q2 vs. Q4: $r=0.228$, $p<0.001$; Q2 vs. Q1: $r=0.078$, $p<0.001$), with effect sizes in the medium range for Q2 vs. Q3 and Q2 vs. Q4. The Q1 stratum also differed significantly from Q3 ($r=0.147$, $p<0.001$) and Q4 ($r=0.129$, $p<0.001$). The sole non-significant comparison was Q3 vs. Q4 ($U=488,020$, $p=0.756$, $r=0.008$), indicating that these two strata are statistically indistinguishable in terms of citation relevance scores. Notably, the effect-size confidence intervals for all significant pairwise contrasts excluded zero, whereas the interval for the non-significant Q3–Q4 contrast included zero (95% CI $[-0.044, 0.058]$), confirming that the practically meaningful differences are confined to comparisons involving Q1 and Q2.

Monotonic trend analysis

Spearman's rank-order correlation between journal quartile rank (Q1 = 1, Q2 = 2, Q3 = 3, Q4 = 4) and citation relevance score yielded $\rho = -0.105$ (95% CI $[-0.132, -0.080]$, $p<0.001$), indicating a statistically significant but weak negative monotonic relationship. This coefficient should not be read as a clean monotonic gradient across the four strata. Because Spearman's correlation operates on score ranks rather than means, the negative sign reflects that the two upper-tier strata (Q1 and Q2) tend to receive higher-ranked relevance scores than the two lower-tier strata (Q3 and Q4): the within-group median score is highest in Q2 (8.90) and Q1 (8.80), and lower in Q3 (8.50) and Q4 (8.20) (see Fig. 4). Q1's rank position is therefore high rather than anomalously low; its lowest-in-dataset arithmetic mean ($M=7.52$) arises because Q1 is bimodal, combining the second-highest proportion of Excellent citations (46.2%) with the highest proportion of Irrelevant citations (10.7%). This bimodal structure — visible in the kernel density estimates (Fig. 1), where Q1 exhibits both a dominant high-score peak and a

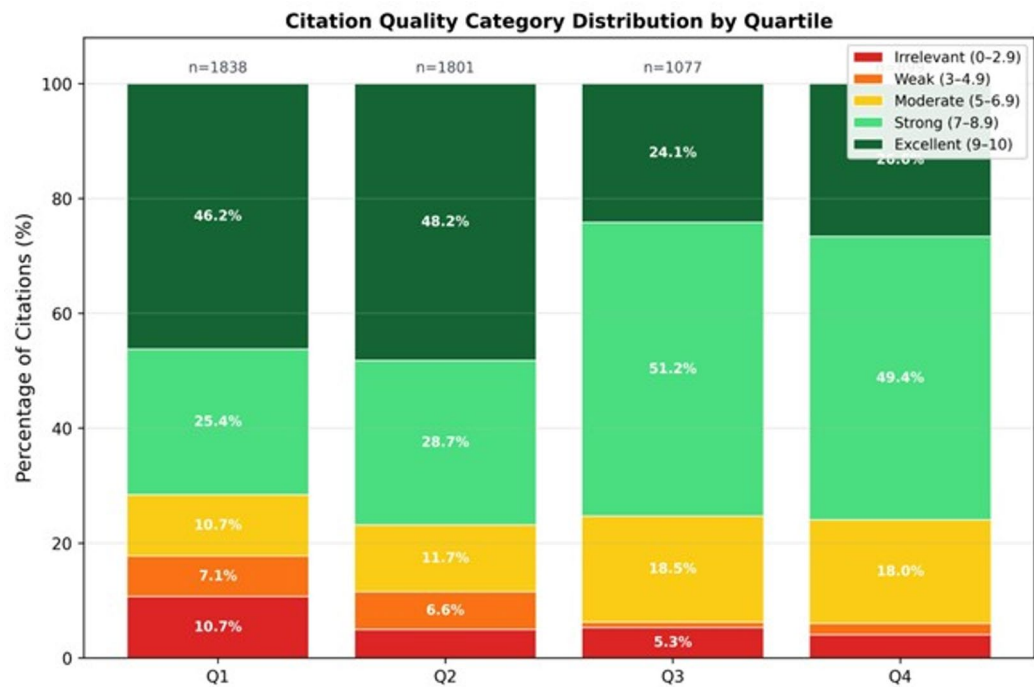


Fig. 3. Citation Quality Category Distribution by Quartile (Stacked Bar Chart). Each bar represents the percentage of citations falling into each quality category within a given quartile stratum.

Comparison	U	p-value	Effect Size (r)	Magnitude	Significant	95% CI (r)
Q1 vs. Q2	1,525,570	4.23×10^{-5}	0.078	Small	Yes	[0.040, 0.116]
Q1 vs. Q3	1,135,394	2.80×10^{-11}	0.147	Medium	Yes	[0.105, 0.190]
Q1 vs. Q4	932,886	3.76×10^{-8}	0.129	Medium	Yes	[0.088, 0.171]
Q2 vs. Q3	1,210,578	4.76×10^{-29}	0.248	Medium	Yes	[0.208, 0.287]
Q2 vs. Q4	993,880	4.23×10^{-22}	0.228	Medium	Yes	[0.185, 0.270]
Q3 vs. Q4	488,020	0.756	0.008	Small	No	[-0.044, 0.058]

Table 5. Pairwise Statistical Comparisons (Mann-Whitney U, Bonferroni Corrected $\alpha=0.0083$). r = rank-biserial correlation; absolute magnitudes are reported, with the direction of each difference indicated by the group means in Table 3. 95% CIs from stratified bootstrap (5,000 resamples).

secondary mass at the low end — depresses Q1’s mean while leaving its median high, so that the weak negative p is consistent with, rather than contradicted by, Q1’s elevated Irrelevant rate. Consistent with the non-significant Q3 vs. Q4 comparison reported above, the two lowest strata are themselves statistically indistinguishable.

Inter-rater reliability validation

To benchmark the LLM-based scoring pipeline against expert human judgment, an inter-rater reliability (IRR) study was conducted on a stratified random subsample of 150 citation–reference pairs (30 per quality category). Three independent expert evaluators, each holding a faculty position in engineering and related computing disciplines, independently classified each pair into one of the five rubric categories (Excellent, Strong, Moderate, Weak, Irrelevant) based on the citation sentence and the referenced article title. Evaluators were blind to the LLM-assigned scores and category labels. The LLM’s continuous scores were mapped to the same five categories using the rubric thresholds defined in Table 1.

Table 6 presents the pairwise agreement statistics. Inter-expert agreement, measured by Cohen’s κ , ranged from 0.287 to 0.343 (Fleiss’ $\kappa = 0.302$, indicating fair agreement under the²⁹ classification). However, Spearman rank-order correlations between expert pairs were substantially higher ($\rho = 0.727\text{--}0.859$, all $p < 0.001$), indicating strong ordinal consistency despite moderate exact-category agreement — a pattern consistent with the inherent subjectivity of fine-grained five-category distinctions in citation relevance assessment. Expert–LLM agreement followed a similar pattern: Cohen’s κ between the expert majority vote and the LLM classification was 0.350, while the corresponding Spearman correlation was $\rho = 0.643$ ($p < 0.001$). Figure 5 visualises the contrast between κ and ρ values across all pairwise comparisons.

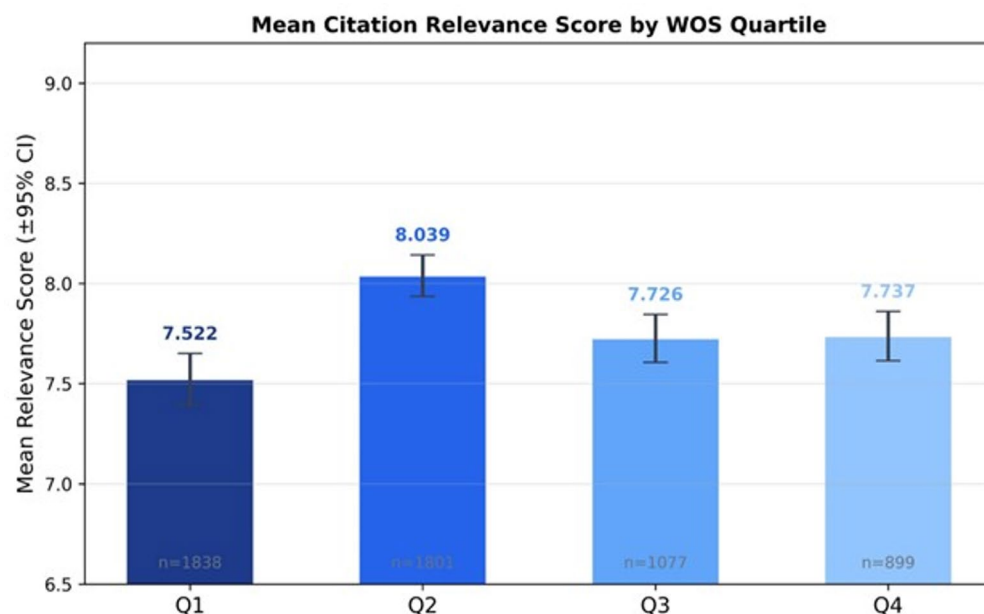


Fig. 4. Mean Citation Relevance Score by WOS Quartile with 95% Confidence Intervals. Error bars represent 95% CI; sample sizes are displayed within each bar.

Comparison	Cohen's κ	Exact Agr.	Spearman ρ	p-value
Expert 1 vs. Expert 2	0.287	41.3%	0.727	< 0.001
Expert 1 vs. Expert 3	0.329	46.7%	0.832	< 0.001
Expert 2 vs. Expert 3	0.343	47.3%	0.859	< 0.001
Expert 1 vs. LLM	0.218	37.3%	0.546	< 0.001
Expert 2 vs. LLM	0.189	35.3%	0.546	< 0.001
Expert 3 vs. LLM	0.350	48.0%	0.782	< 0.001
Majority Vote vs. LLM	0.350	48.0%	0.643	< 0.001
Fleiss' κ (3 Experts)	0.302			

Table 6. Inter-Rater Agreement Between Expert Evaluators and LLM Pipeline. Fleiss' κ interpretation: Fair agreement²⁹. All Spearman correlations significant at $p < 0.001$.

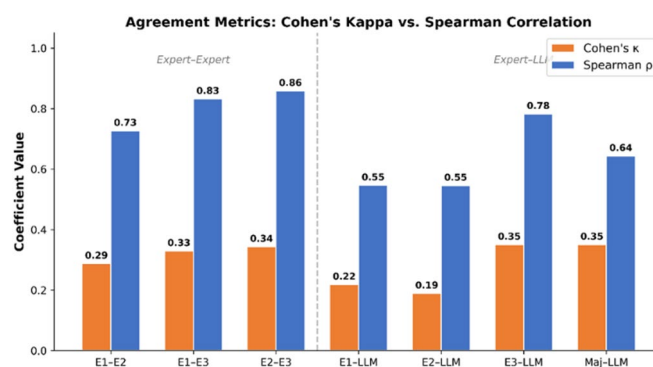


Fig. 5. Agreement Metrics: Cohen's Kappa vs. Spearman Correlation.

Table 7; Fig. 6 report the per-class agreement between the expert majority vote and the LLM classification, computed within each LLM-assigned category. The pipeline achieved its highest exact agreement on the Irrelevant (80.0%) and Excellent (71.0%) categories — the two poles of the scoring rubric — while agreement was lower for the intermediate categories of Weak (36.7%), Strong (37.9%), and Moderate (13.3%). When a ± 1

Category	N (LLM)	Exact Match	%	± 1 Adjacent	%
Irrelevant	30	24/30	80.0%	25/30	83.3%
Weak	30	11/30	36.7%	22/30	73.3%
Moderate	30	4/30	13.3%	19/30	63.3%
Strong	29	11/29	37.9%	23/29	79.3%
Excellent	31	22/31	71.0%	27/31	87.1%
Overall	150	72/150	48.0%	116/150	77.3%

Table 7. Per-Class Agreement Between Expert Majority Vote and LLM Classification (N=150).

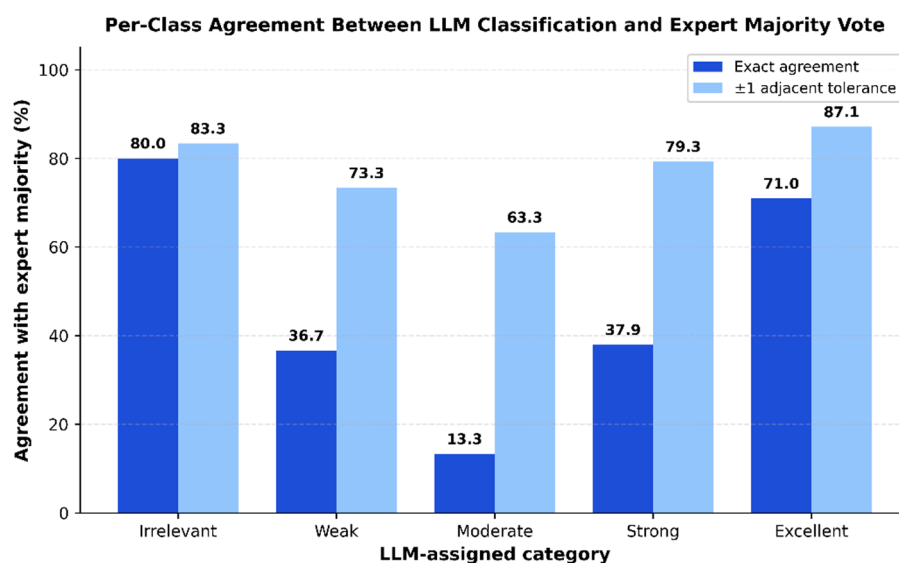


Fig. 6. Per-class agreement between the LLM classification and the expert majority vote. Bars show exact-category agreement and agreement under a ± 1 adjacent-category tolerance, computed within each LLM-assigned category ($n = 30$ per category, except Strong $n = 29$ and Excellent $n = 31$).

LLM $\rightarrow$ Majority $\downarrow$	Irrelevant	Weak	Moderate	Strong	Excellent	Total
Irrelevant	24	8	9	3	2	46
Weak	1	11	7	3	0	22
Moderate	1	3	4	5	2	15
Strong	4	6	8	11	5	34
Excellent	0	2	2	7	22	33

Table 8. Confusion Matrix — Expert Majority Vote vs. LLM Classification. Diagonal cells (shaded) indicate exact agreement. Rows represent expert majority classification; columns represent LLM classification.

adjacent-category tolerance was applied, overall agreement increased from 48.0% to 77.3%, with per-class rates ranging from 63.3% (Moderate) to 87.1% (Excellent).

Table 8; Fig. 7 present the full confusion matrix. The concentration of observations along and near the diagonal confirms that disagreements between the LLM and expert majority were predominantly limited to adjacent categories rather than reflecting gross misclassifications.

To complement the quantitative agreement metrics and to clarify the pipeline's strengths and limitations, we examined the disagreement cases qualitatively. Of the 150 pairs, 78 showed a category mismatch between the expert majority vote and the LLM, of which 44 were adjacent-category disagreements and 34 spanned two or more categories; the LLM assigned a higher category than the experts in 44 cases and a lower category in 34. Table 9 presents four representative disagreements. In two, the LLM over-credited the citation: in case A, the citing sentence attributed a SIFT-based method to a reference describing a biologically inspired, appearance-based SLAM system — a domain-specific misattribution that all three experts flagged as Irrelevant but that the LLM, detecting topical overlap, rated Moderate; in case B, a reference cited only as the source of benchmark problem

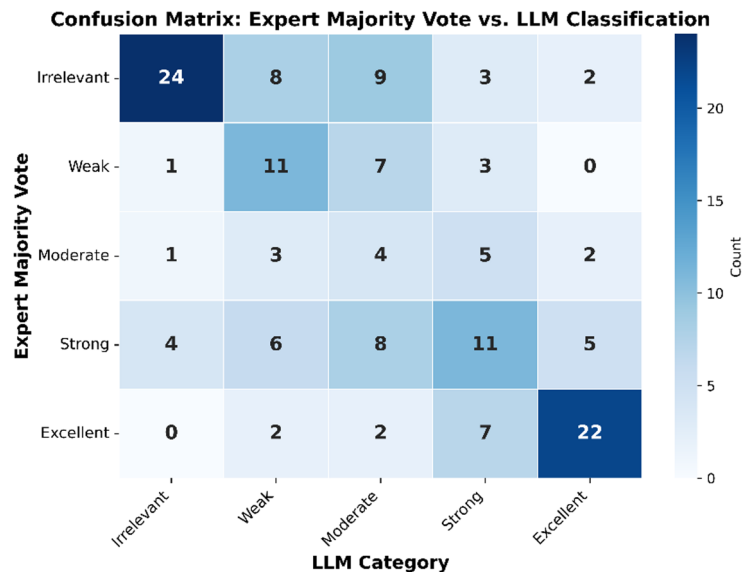


Fig. 7. Confusion matrix of expert majority vote versus LLM classification across the five rubric categories. Cell values are counts; rows represent the expert majority classification and columns the LLM classification; diagonal cells (outlined) indicate exact agreement.

Case	Citing sentence (abbreviated)	Cited reference (short)	Expert majority	LLM (score)	Likely reason
A	“...proposed a SIFT-based corresponded method.”	Biologically inspired SLAM (RatSLAM)	Irrelevant (3/3)	Moderate (6.5)	Topical overlap masks factual misattribution
B	“...we selected nine instances from [...]” (kro-AB100, ...)	Multiobjective EA/SPEA benchmark	Irrelevant (2/3)	Strong (8.5)	Benchmark/dataset-source citation over-credited
C	“...not guaranteed to contain any semantically relevant information...”	Context Encoders (feature learning)	Excellent (2/3)	Weak (3.5)	Contrastive framing under-scored
D	“...deep learning-based single image SR was suggested by...”	Down-sampling video coding/restoration net	Strong (2/3)	Irrelevant (2.1)	Bare numeric reference unresolved

Table 9. Representative Expert–LLM Disagreement Cases.

instances was treated as peripheral by the experts but rated Strong by the LLM, which matched the named instances to the cited article. In the other two, the LLM under-credited the citation: in case C, a contrastively framed sentence engaging the cited work’s core contribution was rated Excellent by the experts but Weak by the LLM, which appears to have read the negative framing as low relevance; and in case D, a citation appearing only as a bracketed numeral, describing a clearly matching super-resolution method, was rated Strong by the experts but Irrelevant by the LLM, consistent with the difficulty of resolving bare numeric references under the model’s skeptical scoring prior. These patterns indicate that the pipeline is most reliable at the extremes of clear relevance and clear irrelevance, and least reliable when relevance hinges on domain-specific factual verification, citation function (e.g., benchmark or dataset citations), contrastive rhetoric, or unresolved reference markers — limitations that motivate the integration of citation-function modelling and reference-resolution checks in future work. More broadly, because these errors fall into a small number of identifiable categories rather than being randomly distributed, several of them may be partly addressable through targeted prompt engineering — for example, explicitly instructing the model to verify the factual attribution of the citing claim, to recognise benchmark or dataset citations as a distinct citation function, and to resolve bracketed numeric references before scoring, consistent with evidence that structured prompting can mitigate systematic biases in LLM judges²⁰. Systematic evaluation of such prompt variants, and of whether they improve agreement without introducing score inflation, is a promising direction for future work.

Sensitivity analysis: pointer citations and robustness of quartile differences
To assess whether the elevated proportion of low-scoring citations in certain quartiles reflects a measurement artifact rather than a genuine bibliographic deficiency, we examined “pointer” citations — formulaic sentences that direct the reader to a source without substantively describing its content (e.g., “Same conclusion can be found in [ref]”; “For the theoretical background, readers can refer to [ref]”). We manually annotated all 378 citations classified as Irrelevant (score < 3.0), labelling each, from the citation sentence alone, as a pointer citation or a genuine semantic mismatch. Overall, 87 of the 378 (23.0%) were pointer citations; their distribution was not uniform across quartiles ($\chi^2(3)=9.74$, $p=0.021$), being highest in Q1 (29.1%) and lowest in Q4 (11.1%) (Table 10).

Quartile	Irrelevant (<i>n</i>)	Pointer <i>n</i> (%)	Genuine mismatch <i>n</i> (%)
Q1	196	57 (29.1%)	139 (70.9%)
Q2	89	14 (15.7%)	75 (84.3%)
Q3	57	12 (21.1%)	45 (78.9%)
Q4	36	4 (11.1%)	32 (88.9%)
Total	378	87 (23.0%)	291 (77.0%)

Table 10. Pointer-Citation Prevalence Among Irrelevant-Classified Citations by WOS Quartile.

Analysis	<i>N</i>	KW <i>H</i>	ϵ^2	Q1 vs. Q2 <i>p</i>	Q1 mean	Q2 mean	Spearman ρ
Baseline (all citations)	5615	157.10	0.028	<0.001	7.52	8.04	−0.105
SA1 (syntactic pointers excluded)	5499	158.73	0.029	<0.001	7.53	8.05	−0.107
SA2 (annotated pointers excluded)	5528	170.66	0.031	0.003	7.73	8.09	−0.126

Table 11. Sensitivity of Quartile Differences to The Exclusion of Pointer Citations. Q1 vs. Q2 *p*-values are from Mann–Whitney U tests (Bonferroni-adjusted $\alpha = 0.0083$). SA1 uses a syntactic pattern filter applied to all citations; SA2 excludes pointer citations identified by manual semantic annotation of the Irrelevant set.

We then tested the robustness of the primary quartile differences to the exclusion of pointer citations using two complementary approaches. First, a syntactic preprocessing filter based on lexical pointer patterns (e.g., “refer to”, “see [“, “for details”) flagged 116 citations (2.1%) across the full dataset; excluding them left all conclusions unchanged (Kruskal–Wallis $H = 158.7$, $p < 0.001$; Q1 vs. Q2 $p < 0.001$). Second, and more substantively, we excluded the 87 manually identified pointer citations and recomputed the primary tests on the remaining 5,528 citations. The omnibus difference remained highly significant ($H = 170.7$, $p < 0.001$, $\epsilon^2 = 0.031$), and the Q1–Q2 contrast — the study’s central finding — remained significant ($p = 0.003$, $r = 0.057$), with Q1 retaining the lowest mean (rising from 7.52 to 7.73) and Q2 the highest (8.09). All other pairwise relationships, including the non-significant Q3–Q4 contrast, were preserved, and the monotonic trend slightly strengthened (Spearman $\rho = -0.126$) (Table 11). Thus, although pointer citations are disproportionately concentrated in Q1 and account for part of its lower mean, the quartile differences are robust and not an artifact of pointer citations.

Discussion

This study proposed and empirically evaluated a two-stage LLM-based pipeline for automated citation quality scoring in Web of Science (WOS)-indexed engineering articles. Across 5,615 scored citation–reference pairs drawn from 121 articles spanning four quartile strata, the pipeline produced scores exhibiting statistically significant variation by journal quartile (Kruskal–Wallis $H(3) = 157.10$, $p < 0.001$), lending preliminary support to the methodological usefulness of the approach as an expert-benchmarked scoring instrument. In the sections that follow, we interpret these findings in relation to the existing literature, address the most notable patterns and anomalies in the data, and situate the present work within the broader context of academic integrity and automated peer review support.

A foundational question underlying this study is whether an LLM can serve as a reliable judge for a domain-specific semantic task — namely, assessing the relevance of a citation sentence to its referenced article. The LLM-as-a-judge paradigm has been systematically validated in general-purpose language evaluation tasks²¹: demonstrated that strong LLM judges such as GPT-4 can achieve agreement rates with human evaluators exceeding 80%, a level on par with inter-human agreement. Thakur et al.²⁰, further documented that LLM judges exhibit systematic biases (including verbosity and position biases) that, while manageable, must be explicitly countered through prompt engineering.

The pipeline incorporated several design features intended to mitigate this tendency. The model was assigned a “skeptical reviewer” persona and given an explicit cautionary instruction about score inflation — an approach consistent with evidence that structured, persona-based prompting can reduce systematic leniency in LLM judges^{20,21}. In addition, the five-level rubric and the requirement of a written justification for each score were intended to discourage inflation by obliging the model to discriminate among explicitly defined qualitative levels and to articulate the basis for each rating. We did not, however, conduct an ablation isolating the contribution of each component, and we therefore present these as design rationales rather than as independently demonstrated effects.

The inter-rater reliability study reported in the section “Inter-rater reliability validation” provides empirical evidence bearing on this question. The observed divergence between Cohen’s κ (fair agreement) and Spearman ρ (strong ordinal correlation) in both expert–expert and expert–LLM comparisons warrants careful interpretation. Cohen’s κ penalises any category mismatch equally, regardless of whether the disagreement is between adjacent categories (e.g., Strong vs. Excellent) or distant ones (e.g., Irrelevant vs. Excellent). For a five-category ordinal scale applied to a subjective semantic judgment, moderate exact-match rates are to be expected even among trained human evaluators. The Spearman correlations — ranging from 0.727 to 0.859 among experts and reaching 0.643 between the expert majority vote and the LLM — indicate that the pipeline’s ordinal ranking of citation quality is substantially aligned with expert judgment, even when exact category boundaries differ. This

pattern mirrors the findings of²⁵, who reported that LLM annotations were comparable to human annotations in consistency but showed lower predictive accuracy relative to gold-standard classifications.

Critically, the pipeline's strongest agreement with experts was observed for the Irrelevant (80.0%) and Excellent (71.0%) categories — the two poles of the quality spectrum that carry the greatest practical significance for editorial screening. A manuscript flagged for a high proportion of Irrelevant citations would warrant closer scrutiny regardless of whether any individual citation was classified as Weak or Moderate by a human reviewer. The lower agreement in intermediate categories (Moderate: 13.3%) reflects the inherent ambiguity of mid-range relevance judgments, which are subject to greater inter-rater variability even among human experts. When a ± 1 adjacent-category tolerance is applied, overall agreement rises to 77.3%, confirming that the pipeline rarely produces grossly discrepant classifications. It should also be noted that the manual evaluation process itself was cognitively demanding: several citation–reference pairs required evaluators to read the referenced article in its entirety to reach a judgment, and evaluator fatigue over successive annotations may have introduced additional variability. This practical constraint — which would equally affect any large-scale human annotation effort — further contextualises the moderate exact-match rates and underscores the value of automated scoring as a complement to, rather than a replacement for, human judgment.

Perhaps the most striking finding in this study is that Q2 articles, rather than Q1 articles, recorded the highest mean citation relevance score (Q2: $M=8.039$; Q1: $M=7.522$). This ordering runs counter to the widely held assumption that citation practices improve monotonically with journal prestige. Several explanations merit consideration.

One plausible structural explanation is that some Q1 engineering articles may make greater use of context-setting citations that situate their contributions within broader interdisciplinary or theoretical frames. In these instances, authors need to place their findings into a large intellectual context which frequently requires authors to cite prior work that connects their current study to related disciplines, research paradigms, etc. as well as methodological approaches (citations) that by definition will be less directly relevant to the main idea/argument being made in the citing sentence. Thus, it can be argued that a lower average citation relevance score for Q1 could reflect an author's greater level of scholarly ambitions, rather than a lack of attention to detail when compiling the reference list.

Second, and more speculatively, the elevated Irrelevant rate in Q1 might in part reflect AI-assisted writing. Q1 journals in the WOS engineering category attract a high volume of submissions worldwide, and recent investigations have documented LLM-hallucinated citations in submissions to top-tier venues — for example, an analysis of papers accepted at NeurIPS 2025 reported more than 100 AI-hallucinated citations across at least 53 papers¹⁰, and a separate investigation identified 50 ICLR 2026 papers each containing at least one hallucinated citation that had passed peer review¹¹. We emphasise, however, that our data provide no direct evidence for this mechanism: the corpus comprises published, human-authored engineering articles, AI involvement was not measured, and the engineering journals studied here differ from the AI-conference settings in which those hallucination rates were observed. This possibility therefore cannot be distinguished, on the basis of our data, from the structural and measurement-related explanations developed above and below, and we present it as a direction for future investigation rather than as a finding of the present study.

This interpretation is reinforced by the pointer-citation analysis reported in the section “[Sensitivity analysis: pointer citations and robustness of quartile differences](#)”. Pointer citations, which carry minimal semantic content in isolation and are therefore prone to low automated scores, were disproportionately concentrated in the Q1 Irrelevant set, indicating that part of Q1's elevated Irrelevant rate is a measurement effect rather than a genuine bibliographic deficiency. Importantly, this effect is only partial: after excluding pointer citations, the Q1–Q2 difference remained significant (Section “[Sensitivity analysis: pointer citations and robustness of quartile differences](#)”), so the quartile pattern cannot be dismissed as an artifact. Taken together with the broader interdisciplinary scope of Q1 articles, these findings support a measurement-plus-structure interpretation of the Q1 pattern rather than any single cause such as AI-assisted writing. Because these citation sentences carry minimal semantic content in isolation, a preprocessing step that detects pointer patterns and either flags them for separate treatment or retrieves additional surrounding context could reduce false positives in the Irrelevant category; the sensitivity analysis in the section “[Sensitivity analysis: pointer citations and robustness of quartile differences](#)” evaluates such an exclusion and confirms that the main findings hold.

It is important to emphasise that the findings reported in this study are not intended to disparage any individual journal, article, author, or indexing service. The quartile-level patterns identified here reflect aggregate statistical tendencies across strata and should not be interpreted as judgments on the quality of specific publications or outlets. The goal of the pipeline is to support — not replace — editorial and peer review processes by providing automated, scalable citation quality indicators.

Third, at a purely descriptive level, Q2 records the lowest combined Irrelevant-plus-Weak citation rate among the two upper-tier strata (Q2: 11.5% versus Q1: 17.7%), although Q3 and Q4 are lower still (6.2% and 6.0%, respectively, reflecting their higher share of Strong rather than Excellent citations). We refrain from attributing this pattern to specific author characteristics — such as seniority, citation diligence, or differing publication pressures — because our data contain no information on author attributes or motivations; the mechanisms underlying the Q2 pattern remain an open question for future research.

The pairwise Mann-Whitney U test for Q3 vs. Q4 returned a non-significant result ($p=0.756$, $r=0.008$), indicating that, in terms of citation relevance, articles published in third- and fourth-quartile journals are statistically indistinguishable. Both groups exhibited comparable mean scores (Q3: 7.726; Q4: 7.737) and

highly similar category distributions, with Strong citations dominant in both groups (51.2% and 49.4%, respectively) and Excellent citations comprising approximately one quarter of each stratum.

This finding has implications for research evaluation policy. Quartile rankings for academic journals within the Journal Citation Reports of Web of Science are inherently based upon a continuum of measurement which can be grouped into distinct ranges - many previous studies have observed that there are few discernible differences in journal performance between adjacent quartiles. Thus, at an empirical level related to citation behavior, the difference between citations to Q3 vs. Q4 publications is likely negligible. That does not mean that the scope, peer-review process or discipline-based coverage of Q3 and Q4 journals would be identical; however, this study suggests that the quartile designation itself is a weak standalone indicator of citation merit among the least prestigious subset of research journals.

Conversely, the significant differences between Q1–Q2 and Q3–Q4 (with Q2 showing the largest effect sizes in pairwise comparisons: $r = 0.248$ vs. Q3 and $r = 0.228$ vs. Q4) indicate that the more meaningful fault line in citation quality lies not between the bottom two quartiles, but between the upper and lower halves of the journal prestige distribution.

LLM-assisted writing has created a new form of error that is both qualitatively different from other forms of bibliography errors (e.g. transcriptions) and quantitatively large. Hallucinated citations are systematically fabricated based on the tendency of language models to create credible-looking, yet unsubstantiated content⁵. What makes these fabricated sources so problematic is that they can appear as plausible, by having authentic-sounding authors; thematic sounding journal titles; and topic-oriented article titles. Consequently, they cannot be reliably identified simply by reviewing the source.

Empirical investigations have documented hallucination rates ranging from approximately 3% to 46% — with an overall rate near 20% — depending on subject domain and prompt specificity⁷. Athaluri⁴, found widespread fabricated references in ChatGPT outputs used for scientific writing, while⁷ reported that citation fabrication rates in mental health research varied substantially with topic familiarity and prompt specificity. Alkaissi and McFarlane³, documented the phenomenon in the context of scientific writing, describing cases in which ChatGPT produced confident but factually incorrect text and entirely fabricated references — a class of error they referred to as ‘artificial hallucinations,’ following the terminology established in the broader NLG literature⁵. In the legal domain, high-profile cases of hallucinated citations — including a case in which fabricated case precedents generated by ChatGPT were submitted in a U.S. federal court filing⁶ — have further underscored the real-world consequences of unchecked LLM outputs.

Current approaches to detecting AI-generated academic content rely primarily on stylometric and linguistic features — statistical properties of text such as perplexity, burstiness, and token predictability — as implemented in tools such as GPTZero, Turnitin AI, and ZeroGPT, whose reliability has been systematically tested¹². While these tools offer scalable screening capabilities, they are known to produce false positives at non-trivial rates, particularly for non-native English speakers, and they remain vulnerable to evasion through paraphrasing^{13,14}.

Citation quality scoring, as proposed in the present pipeline, offers a structurally different and complementary mode of verification. Whereas stylometric detectors analyze the surface properties of text, citation quality scoring probes the semantic relationship between a claim and its evidential basis. When the cited document exists and is available, a fabricated or semantically mismatched citation would be expected to receive a low relevance score, because the referenced content does not support the citing claim; non-existent references, however, require a separate existence-verification step before relevance scoring, as noted in the Limitations. More broadly, text produced through wholesale LLM generation tends to rely on generic, high-level citations that serve a rhetorical rather than a substantive function, a pattern that would be captured as a higher proportion of low-relevance citations by the present scoring rubric.

Citation quality scoring could also serve as an indirect signal when screening for anomalous bibliographic practices. A manuscript containing many low-relevance citations — concentrated in the Irrelevant (0.0–2.9) or Weak (3.0–4.9) bands — could be flagged for editorial review: not as conclusive evidence of misconduct, but as a quantifiable indicator that human scrutiny is warranted. Consistent with this view, citation-quality validation has been incorporated into emerging AI-assisted paper-mill detection systems used by several major publishers³⁰. We emphasise, however, that our study provides no evidence linking any particular quartile to paper-mill activity; the elevated Irrelevant rate in Q1 is, as discussed above, better explained by interdisciplinary scope and a pointer-citation measurement effect.

It is important to note that the converse is not necessarily true: a paper with uniformly high citation scores may still be entirely AI-generated if the model had access to the actual referenced documents (e.g., through retrieval-augmented generation). Citation quality scoring is therefore most powerful as one layer within a multi-modal integrity framework, combined with stylometric, statistical, and human review components.

A central premise underlying the practical utility of automated citation quality scoring is that manual verification of citations by peer reviewers is, under current conditions, structurally infeasible. This is not merely a matter of reviewer diligence; it reflects fundamental constraints in the peer review system.

The peer review system operates under severe and worsening resource constraints. Submission volumes have grown dramatically — in 2022 alone, approximately 2.82 million journal articles were indexed in Scopus and Web of Science — a 47% increase compared to 2016¹ — while the pool of active reviewers has remained stagnant or contracted. Peer review is unpaid and, in most institutional contexts, unrecognized in

workload models². Surveys consistently identify lack of relevant expertise and time pressure as the primary reasons for declining review invitations³¹, with review delays disproportionately harming early career researchers whose career progression depends on timely publication decisions. Under these conditions, the cognitive budget available to a typical reviewer is sufficient for evaluating the intellectual contribution, methodological soundness, and logical coherence of a manuscript — not for systematically locating, accessing, and reading every cited document in a reference list that may contain 40, 60, or more entries.

The practical challenge is compounded by access barriers. Many referenced articles are not openly available, requiring institutional subscriptions or author-side access that a reviewer cannot be assumed to have. Even for open-access papers, the time required to locate, download, and read a referenced article to verify a single citation sentence may exceed the entire time a reviewer allocates to the manuscript. The consequence is that citation lists routinely pass through peer review without substantive verification — a structural vulnerability that predates, but is now substantially amplified by, the rise of LLM-assisted writing.

Automated citation quality scoring directly addresses this gap. A pipeline of the kind proposed here can process an entire reference list in minutes, flagging citations whose relevance score falls below a defined threshold for targeted human review. Rather than replacing peer review, the pipeline functions as a pre-screening layer that directs reviewer attention to the citations most likely to be problematic, effectively converting an infeasible task (checking all citations) into a tractable one (verifying a small, flagged subset). This division of labor — automated breadth, human depth — is designed to preserve human oversight, particularly given documented concerns about LLM-generated reviews lacking genuine critical content²².

Limitations

Several limitations of the present study should be acknowledged. First, the pipeline scores the semantic relationship between a citation sentence and a referenced article, but cannot distinguish between three distinct low-score scenarios: a genuine academic article that is simply irrelevant to the citing claim; a hallucinated article that does not exist; and a genuine article that was inaccessible to the scoring model due to access restrictions. In the present dataset, 1,015 citation pairs (12.1%) received no score due to full-text inaccessibility, and four citation instances in the Q4 stratum pointed to non-existent sources and were assigned scores of 0. While these anomalies were handled transparently, the current system conflates genuine irrelevance with fabrication, which limits its utility as a standalone hallucination detector. Future implementations should integrate an existence-verification step (e.g., querying DOI registries and academic metadata APIs) prior to relevance scoring, enabling the system to distinguish “cannot retrieve” from “does not exist” from “exists but irrelevant.” Moreover, the pipeline scores citation sentences in isolation, which disadvantages “pointer citations” — formulaic sentences that direct the reader to a source without describing its content — yielding low scores despite a genuine underlying relationship.

Because generation parameters were left at their default values and no random seed was fixed, the pipeline's outputs are not strictly deterministic; identical inputs may yield slightly different scores across runs. The large sample size mitigates the effect of this variation on aggregate quartile-level statistics, but exact reproduction of individual scores is not guaranteed. A further limitation concerns the inter-rater reliability study. The expert evaluators classified each citation using the citation sentence and the referenced article's title, whereas the Stage 2 model scored each citation using the citation sentence and the reference's full text; the two raters therefore operated on different information bases. This asymmetry is most likely to deflate, rather than inflate, the reported expert–LLM agreement: when a title is generic or uninformative, an evaluator restricted to the title cannot recover the full-text relevance signal available to the model, so genuine agreement at the full-text level may register as disagreement. Consistent with a conservative rather than inflated estimate, the expert-majority-versus-LLM exact-category agreement ($\kappa=0.350$) was comparable to, and slightly higher than, the agreement among the experts themselves (Fleiss' $\kappa=0.302$), despite the model's information advantage. We therefore interpret the expert–LLM comparison as a conservative benchmark of the model against rapid, title-based human judgment rather than as a strict test of convergent validity, and we note that a fuller-information replication — providing evaluators with reference abstracts or full texts — is a valuable direction for future work.

Second, the study is restricted to the engineering domain within WOS. Citation norms vary substantially across disciplines — the function and style of citations in computational engineering, for example, differ markedly from those in theoretical physics, biomedicine, or the social sciences. The generalizability of the quartile patterns observed here to other disciplinary contexts cannot be assumed and requires independent empirical evaluation. A further limitation concerns the construct measured by the rubric. The pipeline scores the semantic centrality of a citation relative to the cited article's content and does not separately model citation function — for instance, whether a citation is methodological (acknowledging a reused protocol or dataset) or substantive. Consequently, a legitimate and appropriate citation to a peripheral element of the cited work may receive an intermediate centrality score despite being entirely appropriate. Future work could combine centrality scoring with citation-function classification^{15–17} to distinguish citations that are peripheral but appropriate from those that are genuinely misaligned.

Third, although the inter-rater reliability study (Section “[Inter-rater reliability validation](#)”) provides initial evidence of agreement between the LLM scores and expert judgment, the validation was conducted on a stratified subsample of 150 pairs drawn from the scored dataset itself, and the expert evaluators classified

citations into five discrete categories rather than assigning continuous scores. The moderate Cohen's κ values (0.29–0.35) suggest that exact-category agreement remains limited, particularly for intermediate quality levels. Future work should extend this validation using larger samples, continuous scoring by experts, and evaluators drawn from diverse engineering sub-disciplines to strengthen the generalizability of the pipeline's scoring.

Fourth, the sampling strategy, while systematic, was constrained by journal availability, article selection criteria, and reference accessibility. The unequal cell sizes across quartiles (Q1: $n = 1,838$; Q4: $n = 899$) reflect these constraints and, while the use of non-parametric statistics mitigates concerns about unequal variance, they limit the statistical power of some comparisons.

Fifth, the pipeline requires access to the full text of both the citing article and every cited reference in order to perform semantic relevance scoring. In practice, this means that authors or editors wishing to use the system must supply not only the manuscript under review but also the full-text PDFs of all referenced works — a requirement that may be constrained by copyright restrictions, institutional access limitations, and the sheer volume of references in a typical manuscript. This dependency on full-text availability represents a significant barrier to deployment in settings where comprehensive library access is not guaranteed, and future implementations should explore abstract-level or metadata-based scoring as a lower-cost alternative for references whose full text cannot be obtained.

Finally, all scores in this study were produced by a single large language model (gemini-2.5-flash), and LLM-as-a-judge evaluations are inherently model-dependent. A different model, a different version of the same model, or even the same model queried at a later date may assign systematically different relevance scores, because the scoring reflects that particular model's learned representations and its calibration to the rubric rather than a model-independent ground truth. The absolute scores, the category cut points, and the precise magnitude of the quartile-level differences reported here should therefore be interpreted as specific to this model and configuration; the core qualitative findings would need to be re-established before being generalized to other scoring models. This concern compounds the run-to-run non-determinism noted in the section “[Methodology](#)”, as generation parameters were left at their defaults and no random seed was fixed. Systematic multi-model validation — comparing scores across several LLMs to assess their cross-model agreement, or aggregating them into an ensemble judgment — is an important direction for strengthening the robustness and transferability of the approach.

Relatedly, although the pipeline includes several features intended to limit score inflation (a skeptical-reviewer persona, an explicit anti-inflation instruction, the five-level rubric, and mandatory justifications), we did not perform an ablation study isolating the contribution of each; their individual effects on score calibration therefore remain to be quantified.

Implications and future directions

The findings of this study carry implications at three levels: methodological, applied, and normative. At the methodological level, the study demonstrates that LLM-based citation quality scoring is feasible at scale and produces scores that exhibit theoretically coherent variation across a journal-level prestige and impact proxy (journal quartile). This establishes a proof-of-concept foundation for more rigorous validation studies that incorporate expert annotation, cross-disciplinary replication, and longitudinal analysis of citation quality trends.

At the applied level, the pipeline — or a refined successor — could be integrated into editorial management systems to serve as an automated pre-screening layer for submitted manuscripts. A citation quality report generated prior to peer review assignment would allow editors to flag manuscripts with anomalous citation profiles for targeted scrutiny, without requiring reviewers to undertake the infeasible task of verifying all references manually. Combined with existing AI-content detectors and statistical integrity tools, citation quality scoring would form a complementary layer in a multi-signal editorial integrity framework.

At the normative level, the study contributes to the emerging debate about the responsibilities of authors, editors, and institutions in the age of AI-assisted writing. The presence of hallucinated citations in peer-reviewed publications — documented at premier venues such as NeurIPS and ICLR and increasingly reported in broader publication contexts by prior work — represents not merely a quality control failure but a fundamental challenge to the verifiability and reproducibility of the scientific record. Citation quality scoring offers one scalable mechanism for surfacing bibliographic anomalies before publication. In this sense, the present pipeline contributes to a broader class of tools — alongside paper-mill detection services and statistical verification software — that may help strengthen scientific integrity in the post-LLM era.

Conclusion

This study proposed, implemented, and evaluated a two-stage LLM-based pipeline for automated citation quality scoring across WOS-indexed engineering journals. Analysis of 5,615 citation–reference pairs revealed statistically significant differences in citation relevance scores across journal quartile strata, with

Q2 articles recording the highest mean scores and Q3 and Q4 articles showing no statistically distinguishable difference. These findings challenge simplistic assumptions about a linear relationship between journal prestige and citation quality. The lower mean score of the Q1 stratum coexists with one of the highest proportions of highly relevant citations and is partly attributable to a disproportionate concentration of context-dependent pointer citations in its Irrelevant set; we therefore read it as a stratum warranting closer methodological attention rather than as evidence of inferior scholarship or of any single cause such as AI-assisted writing.

More broadly, this study argues that citation quality scoring can serve as a valuable complement to existing academic integrity mechanisms. As peer review labor becomes increasingly scarce and AI-generated content becomes increasingly prevalent, automated tools that can assess the semantic integrity of citations at scale — and thereby surface potential bibliographic anomalies for human verification — may provide an increasingly useful complementary layer. The pipeline presented here represents a first step toward such a system. An inter-rater reliability study demonstrated that the pipeline achieves strong ordinal agreement with expert evaluators (Spearman $\rho = 0.643$) and high exact-match rates at the critical Irrelevant (80.0%) and Excellent (71.0%) poles, supporting its practical applicability for editorial pre-screening. Further validation with larger expert panels and integration into editorial workflows constitute the most important directions for future work.

Data availability

De-identified data sufficient to reproduce the primary statistical analyses reported in this article are openly available in the Kaggle repository (<https://doi.org/10.34740/kaggle/ds/10865061>). The deposited files contain only anonymized identifiers, the numeric relevance scores, the ordinal expert labels, the manual pointer labels, and the quartile strata. The one analysis that cannot be reproduced from these files is the syntactic pointer-filter analysis (SA1), which operates on the verbatim citation text. Because the corpus comprises third-party copyrighted articles, the full verbatim citation-sentence dataset and the identities of the citing and cited articles are not redistributed. The only text shared for transparency is a small number of selected illustrative citation excerpts presented in the manuscript and the repository, each accompanied by a neutral content summary of the cited reference rather than its title. Researchers can regenerate the text-level data by applying the openly available pipeline to their own document corpus.

Code availability

The full two-stage pipeline and the exact prompts are openly available in the Kaggle repository (DOI: <https://doi.org/10.34740/kaggle/ds/10865061>).

Received: 17 April 2026; Accepted: 30 June 2026

Published online: 03 July 2026

References

- Hanson, M. A., Barreiro, P. G., Crosetto, P. & Brockington, D. The strain on scientific publishing. *Quant. Sci. Stud.* **5** (4), 823–843 (2024).
- Jamali, H. R., Luca, E. & Wakeling, S. The peer review system is breaking down. Here's how we can fix it. *The Conversation* (2026).
- Alkaissi, H. & McFarlane, S. I. Artificial hallucinations in ChatGPT: implications in scientific writing. *Cureus*, **15**(2). (2023).
- Athaluri, S. A. et al. Exploring the boundaries of reality: Investigating the phenomenon of artificial intelligence hallucination in scientific writing through ChatGPT references. *Cureus* <https://doi.org/10.7759/cureus.37432> (2023).
- Ji, Z. et al. Survey of hallucination in natural language generation. *ACM Comput. Surveys.* **55** (12), 1–38 (2023).
- Neumeister, L. Lawyers blame ChatGPT for tricking them into citing bogus case law. *AP News* 7 (2023).
- Linardon, J. et al. Influence of topic familiarity and prompt specificity on citation fabrication in mental health research using large language models: Experimental study. *JMIR Ment. Health* **12**, e80371 (2025).
- Mugaanyi, J., Cai, L., Cheng, S., Lu, C. & Huang, J. Evaluation of large language model performance and reliability for citations and references in scholarly writing: Cross-disciplinary study. *J. Med. Internet Res.* **26**, e52935 (2024).
- Xu, Z. et al. GhostCite: A Large-Scale Analysis of Citation Validity in the Age of Large Language Models. arXiv preprint arXiv:2602.06718 (2026).
- Goldman, S. 2026/01/21. NeurIPS research papers contained 100+ AI-hallucinated citations, new report claims. (2026). <https://fortune.com/2026/01/21/neurips-ai-conferences-research-papers-hallucinations/>
- Martin, S. *Startup investigation reveals 50 peer-reviewed papers contained AI-hallucinated citations*. Retrieved 10.03.2026 from (2025). <https://betakit.com/start-up-investigation-reveals-50-peer-reviewed-papers-contained-hallucinated-citations/>
- Weber-Wulff, D. et al. Testing of detection tools for AI-generated text. *Int. J. Educ. Integr.* **19**(1), 1–39 (2023).
- Liang, W., Yuksekgonul, M., Mao, Y., Wu, E. & Zou, J. GPT detectors are biased against non-native English writers. *Patterns* **4**(7) (2023).
- Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W. & Feizi, S. Can AI-generated text be reliably detected? arXiv preprint arXiv:2303.11156 (2023).
- Wan, X. & Liu, F. Are all literature citations equally important? Automatic citation strength estimation and its applications. *J. Assoc. Inform. Sci. Technol.* **65**(9), 1929–1938 (2014).
- An, X., Sun, X. & Xu, S. Important citations identification with semi-supervised classification model. *Scientometrics* **127** (11), 6533–6555 (2022).
- Zhang, Y. et al. Deep learning meets bibliometrics: A survey of citation function classification. *J. Informetrics.* **19** (1), 101608 (2025).
- Haan, S. SemanticCite: Citation Verification with AI-Powered Full-Text Analysis and Evidence-Based Reasoning. arXiv preprint arXiv:2511.16198 (2025).
- Wu, K. et al. An automated framework for assessing how well LLMs cite relevant medical references. *Nat. Commun.* **16**(1), 3615. <https://doi.org/10.1038/s41467-025-58551-6> (2025).
- Thakur, A. S., Choudhary, K., Ramayapally, V. S., Vaidyanathan, S. & Hupkes, D. Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)* (2025).
- Zheng, L. et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Adv. Neural. Inf. Process. Syst.* **36**, 46595–46623 (2023).

22. Donker, T. The dangers of using large language models for peer review. *Lancet. Infect. Dis.* **23** (7), 781 (2023).
23. Iqbal, S. et al. A decade of in-text citation analysis based on natural language processing and machine learning techniques: an overview of empirical studies. *Scientometrics* **126** (8), 6551–6599 (2021).
24. Jiang, X. & Chen, J. Contextualised segment-wise citation function classification. *Scientometrics* **128**(9), 5117–5158 (2023).
25. Nishikawa, K. & Koshiba, H. Exploring the applicability of large language models to citation context analysis. *Scientometrics* **129** (11), 6751–6777 (2024).
26. Jurgens, D., Kumar, S., Hoover, R., McFarland, D. & Jurafsky, D. Measuring the evolution of a scientific field through citation frames. *Trans. Assoc. Comput. Linguist.* **6**, 391–406 (2018).
27. Teufel, S., Siddharthan, A. & Tidhar, D. Automatic classification of citation function. In *Proceedings of the 2006 conference on empirical methods in natural language processing* (2006).
28. Zhuang, Z., Chen, J., Xu, H., Jiang, Y. & Lin, J. Large language models for automated scholarly paper review: A survey. *Inf. Fusion* **124**, 103332 (2025).
29. Landis, J. R. & Koch, G. G. The measurement of observer agreement for categorical data. *Biometrics* **33**(1), 159–174. <https://doi.org/10.2307/2529310> (1977).
30. Pellegrina, D. & Helmy, M. AI for scientific integrity: Detecting ethical breaches, errors, and misconduct in manuscripts. *Front. Artif. Intell.* **8**, 1644098 (2025).
31. Felisi, N., Vecchio, R. & Odone, A. When peer review drags on: The harm to early career researchers. *Front. Res. Metr. Anal.* **11**, 1740381 (2026).

Acknowledgements

The citation–reference pair dataset generated and analysed during this study cannot be made publicly available. Reproducing the dataset requires access to the full-text PDFs of both the citing and cited articles, the majority of which are protected by publisher copyright and accessible only through institutional subscription agreements.

Author contributions

M.I. conceptualized the study, designed the methodology, and substantially revised and finalized the data analysis and manuscript. S.G. developed the automated LLM-based pipeline, conducted data collection and statistical analysis, and contributed to drafting the manuscript. M.I. supervised the research process and led the interpretation of the results. All authors reviewed, verified, and approved the final manuscript.

Funding

This research received no external funding.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to M.I.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2026