

ReCitE: Evaluating Citation Context Quality and Relevance with Large Language Models

Ahmar K. Hussain^{1*}, Bernhard A. Sabel² and Andreas Nürnberger¹

¹Otto von Guericke University Magdeburg, Germany

²Sciii Foundation gGmbH, Berlin, Germany

Abstract

The growing number of fraudulent publications poses a serious threat to scientific integrity; however, a substantial number of these publications remain undetected in the published literature. Irrelevant citations, among other integrity problems, is a prevalent concern in fraudulent papers. To help address this problem, we developed ReCitE (Reliable Citation Evaluation): a framework to identify fraudulent papers by analyzing citation relevance, that is, whether a reference actually supports the claim to which it is attached. The framework includes parsing a PDF, extracting references, finding the referenced papers in open bibliographic databases, then prompting a large language model to identify the claims tied to each citation, score their relevance, and provide a brief justification for each assessment. We manually evaluate the quality of citation context extraction and evaluate citation relevance with five open-weight, locally deployable LLMs on a public dataset of 22,381 labeled citations. The best-performing model for citation relevance was Gemma 2-27b, with an F1 score of 0.981 and an ROC AUC of 0.981, comparable to the state-of-the-art results based on SBERT embeddings. An error analysis shows that misclassifications occur due to generic citation contexts, uninformative abstracts and citations used to establish background rather than support a direct claim. This method is not intended to replace human reviewers, but rather to serve as an automated tool that helps reviewers, editors and integrity sleuths flag papers for closer inspection.

Keywords

Scientific misconduct, Citation analysis, Research integrity, Bibliometrics, Large language models

1. Introduction

The integrity of scientific literature has come under increasing scrutiny as the scale and sophistication of fraud in academic publishing has grown [1]. Among the most damaging actors in this publishing ecosystem are so-called "paper mills", commercial entities that produce fabricated publications in exchange for monetary gain. Paper-mill-produced papers have infiltrated the scientific record, inflated author citation metrics, corrupted the scientific literature, and eroded public trust in science [2, 3]. A growing number of such fraudulent papers have been documented in the Retraction Watch database [4] and RetractableBase [5], yet a substantial proportion of paper mill output remains undetected in the literature [6].

Detecting fraudulent papers is a multifaceted challenge. Previous studies have focused on detecting anomalies in the content of a paper, such as "tortured phrases" [7], or issues with metadata [1], while other approaches have focused on using machine learning methods to classify fake papers [8, 9, 10]. However, one area that has received less attention is the analysis of citation behavior as a potential tool for identifying fraudulent papers. One form of citation manipulation is the insertion of references that are not relevant to the text in which they are cited [11]. These types of citation manipulations constitute a semantic problem in academic papers; therefore, we propose a method for detecting them.

Papers containing irrelevant references present a problem that requires a semantic understanding of both the referenced paper and the text making the claim to determine whether the citation is justified. This task is well suited to large language models (LLMs), given their ability to process semantic meaning of text instead of relying on simple keyword matching.

6th International Workshop on Scientific Knowledge: Representation, Discovery, and Assessment, Oct 2026, Bari, Italy

*Corresponding author.

✉ ahmar.hussain@ovgu.de (A. K. Hussain)

🆔 0009-0002-8223-4638 (A. K. Hussain); 0000-0002-4472-5543 (B. A. Sabel); 0000-0003-4311-0624 (A. Nürnberger)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

This study presents a citation relevance analysis framework called ReCitE (Reliable Citation Evaluation), in which a research paper is first parsed to extract its references and other relevant information. The extracted data are then provided to an LLM, which assesses the relevance of each citation to the claims made in the paper. We evaluate the quality of the context extraction manually and the context relevance using five LLMs of varying sizes on a publicly available labeled dataset [12].

Our approach constitutes a fully automated pipeline for citation relevance analysis from parsing a research paper in PDF format to scoring the relevance of each of its references. It demonstrates that citation relevance issues provide a meaningful signal for identifying fraudulent papers. It should be noted that the presence of irrelevant citations represents only one of several signals that may indicate a potentially fraudulent paper. Therefore, a paper containing irrelevant citations should not, by itself, be classified as fraudulent. Rather, such citations should serve as a flag warranting further investigation of the paper and consideration of additional indicators. We discuss the details of the study, the models' performance, and their limitations in the following sections.

2. Related work

Only a handful of studies have addressed the problem of irrelevant citations directly. The task requires a semantic understanding of both the citing text and the cited work, making it well suited for LLMs.

Citelens [13] is an early industrial tool for this purpose. It uses a Specter academic language model to assess the (1) relevance between an article and its references, and (2) relevance between references and their citation context using cosine similarity as an evaluation metric. Although CiteLens is already integrated into editorial workflows, it cannot identify the specific claims being made in a paragraph containing citations.

Another, more recent study, CiteScreener [12], offers an open-source pipeline, including citation context extraction, abstract retrieval, and a multi-method citation relevance evaluation using both LLMs and text overlap measures. One limitation of this study is the extraction of the claim related to a citation in the text. It is extracted using a fixed character window approach, which may not extract the exact claim in every case. Some citations may refer to claims made several sentences earlier, while others may refer to only a few preceding words, so a fixed window may either over- or under-capture the context depending on the case. Therefore, we hypothesize that extracting context based on the semantic meaning of the text, rather than using a fixed span would provide better results. The citations in this approach are assigned a binary label i.e. either "Related" or "Unrelated". While this approach is straightforward, a binary classification loses nuance that is valuable in practice. Their best citation relevance results on a dataset with 23,308 citations achieved an F1 score of 0.985 and an accuracy of 0.981, using an SBERT model to produce embeddings.

A parallel line of work approaches citation integrity from a knowledge graph perspective. Huang et al. [14] propose CE-KG, a framework that develops a knowledge graph linking claims from the conclusions of research papers to citation contexts using sentiment polarity as a proxy for the citing community's evaluation of a claim. Unlike CiteLens and CiteScreener, which evaluate the relevance of individual citations to their local claims in a paper, CE-KG aggregates citation evidence from a knowledge graph to support the credibility tracing of knowledge claims.

Taken together, the studies discussed in this section suggest that semantic-based detection of integrity issues is a feasible approach for identifying potentially fraudulent or paper-mill-generated papers. Our work builds on this foundation by refining and automating the process of citation context extraction and by providing an explicit justification for each relevance score, offering both improved precision of citation context extraction and greater interpretability than prior approaches.

3. Citation relevance analysis

We examine citation relevance at the individual citation level and compare multiple large language models to assess whether each citation supports the claim being made in the respective paper.

3.1. Motivation

Irrelevant citations are a well-documented problem in biomedical and life science literature [15], motivating the need for a pipeline that could automatically detect and extract them. We carry out this analysis to determine whether a given reference in a paper is actually meaningful and related to the claims made in the paper. This is done by checking if the claim made in a paper is supported by the abstract of the referenced paper. If multiple references are not relevant to the claims in the paper, the paper was flagged for integrity concerns.

To perform this assessment, we use an LLM to evaluate the relevance of a citation context based on the abstract of the cited paper. We decided to use LLMs for this task due to their ability to process the semantic meaning of text. The pipeline from PDF parsing to scoring the relevance is depicted in Fig. 1.

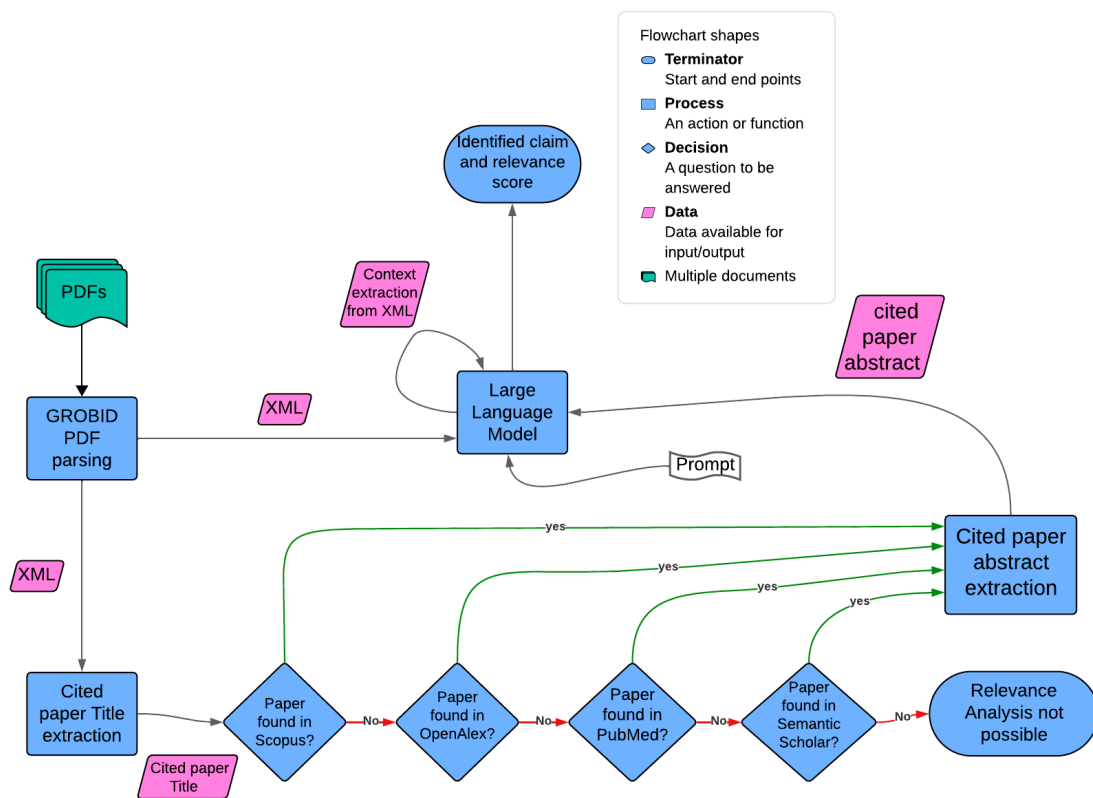


Figure 1: Workflow of citation relevance analysis

3.2. PDF parsing

To analyze the content of a given PDF, we used GROBID (GeneRation Of Bibliographic Data) [16], as our PDF parsing tool to convert the paper into a machine-readable format. GROBID is an open-source tool that restructures a raw PDF into structured XML/TEI document. It extracts the title, authors, affiliations, abstract, citations, and other bibliographic information. Because it also recognizes the different sections and paragraphs of the paper, it is useful for identifying the claim related to a given citation.

While GROBID is the most popular open-source tool for PDF parsing of academic documents, it is not without limitations. In particular, we observed cases where GROBID misidentifies equation numbers as citations. We exclude such cases from our analysis and note this as a limitation of the approach. GROBID is deployed via Docker and we used its "Process Full Text" functionality to parse each PDF in full, returning a structured XML file used for downstream analysis.

3.3. Large Language models

We use a large language model for three tasks in this study:

1. Identifying the claim associated with a citation.
2. Judging the relevance of the citation to that claim.
3. Providing a brief justification for the relevance score.

Given that LLMs have already been trained on vast amounts of data and have the ability to process natural language, we leverage their ability to make the final judgment on relevance. As shown in Fig. 1, the LLM is provided with the abstract of the cited paper and the XML document of the main paper, from which it extracts the claims associated with each citation. A citation may correspond to a single claim or multiple claims, since it can appear at multiple locations throughout the paper; when a citation appears more than once, a relevance score is assigned to each instance separately.

Additionally, context extraction is handled by the LLM itself, in contrast to the previous study [12] discussed in the literature review, where the claim is defined as a number of characters before and after a citation. We assign this task to the LLM by providing it with the paragraph containing the citation and prompting it to identify the claim associated with the citation. The full prompt given to the LLM is provided in Appendix A. The LLM’s output consists of the identified claim extracted from the context, a relevance score between 1 and 10, and a brief justification for the decision.

3.4. Cited paper extraction

Since GROBID identifies the citations within the XML, the titles of the cited papers are used to search for them in open bibliographic databases. As shown in Fig. 1, each title is queried against Scopus, OpenAlex, PubMed, and Semantic Scholar, using their respective APIs. If the paper is not found in these databases, citation relevance analysis cannot be performed, which is rarely the case. When a paper is found, its abstract is extracted and provided to the LLM for analysis.

We chose the paper title as a search query because, for most papers, it is almost always available in the reference list, unlike the DOI, which would have provided a more robust method, but is not consistently available in the source PDF.

This approach has some drawbacks. In particular, the parser does not always correctly extract the title; for example, titles beginning with a number often have that number omitted, causing the corresponding paper to not be retrieved.

3.5. Which LLMs are used and why?

We exclusively use open-weight and locally deployable models, allowing the study to be reproduced without dependence on commercial LLMs, e.g., ChatGPT and Claude. We evaluate citation relevance using open-weight models: Qwen 2.5-7b and Mistral-7b, as they have been used by a previous study [12], and we use additional ones, including larger models, to investigate whether they provide improved performance. We use Gemma 2-9b, Gemma 2-27b, and Llama 3.2-3b. These models were selected to represent a range of parameter sizes in order to assess whether the size of the model would affect the performance for this specific task. All models were run locally using Ollama on a virtual machine with an NVIDIA RTX 6000 with 48GB of VRAM. The models were evaluated using accuracy, precision, recall and F1 scores. The results of the evaluation of these models are provided in Table 2. Gemma 2-27b provides the best results in almost all evaluation metrics.

3.6. Relevance score

A previous study [12] used binary labels to describe the relevance of a cited article to its context (‘Related’ or ‘Unrelated’). As discussed in the literature review section, this makes the classification lose nuance: a citation that is somewhat related to the general area of the claim is different from a completely irrelevant one, yet both are classified as irrelevant under a binary scheme. Our approach, therefore,

assigns citations a score between 1 and 10, where 1 indicates that the citation is completely irrelevant, and 10 indicates that the citation is directly supported by the claim in the paper. This granularity and the justification provided by the model allows the reviewer to distinguish between intentional fake citations, and merely weak or contextual citations, which is important to avoid false accusations against authors who cite a paper for its general idea rather than a specific claim. In addition to the score, the LLM is also tasked to provide a brief justification of its judgment giving a human reviewer better insight into why a citation received a particular score when checking a manuscript .

For the purpose of evaluation, we set a threshold in our scoring system: citations scoring above this threshold are classified as 'relevant', and all others classified as 'irrelevant'. The threshold is selected on the basis of optimization of the F1 score, reflecting a tradeoff between precision and recall.

We also note the possibility of falsely flagging legitimate citations as illegitimate. As such, the output of this approach should always be reviewed by a human and not treated as a final verdict.

4. Evaluation

4.1. Evaluation of context extraction

One of the main contributions of this study is the automated extraction of the citation contexts from papers. The previous study [12] obtained contexts from PubMed Baseline XML, which is limited to open-access papers. A limitation of this approach is that it can only be applied to already published open-access papers. Our approach, by contrast, extracts context directly from PDFs, allowing it to be applied to both unpublished manuscripts and published papers.

To evaluate the quality of the extracted context, we manually assessed 100 randomly sampled citations in the dataset provided by CiteScreener. This evaluation was conducted manually because no ground truth exists for citation context quality in the dataset. Therefore, human judgment served as the basis for assessing accuracy. The annotation was done by the lead author (Ahmar K. Hussain) as a blind evaluation, i.e., the annotator did not know beforehand whether a given context originated from CiteScreener or from ReCitE.

The context quality evaluation is divided into four categories as shown in Table 1. It can be seen that for around 51% of the cases our approach had a more precise extraction of the context than the previous study. For 36% of the cases the context was identical and accurate. Additionally, detailed examples of all four categories can be seen in Appendix B

Table 1

Comparison of context extraction quality between ReCitE and CiteScreener

Category	Explanation	Count
ReCitE is more accurate	ReCitE context includes the full claim; CiteScreener's is incomplete, truncated, or includes irrelevant text	51
Both are accurate	Both approaches extract functionally equivalent context capturing the same claim	36
ReCitE is less accurate	ReCitE context is accurate but omits part of the relevant claim that CiteScreener captured	10
ReCitE is incorrect	ReCitE context does not correspond to the actual claim (wrong sentence/location, or extraction failure)	3

4.2. Evaluation of context relevance

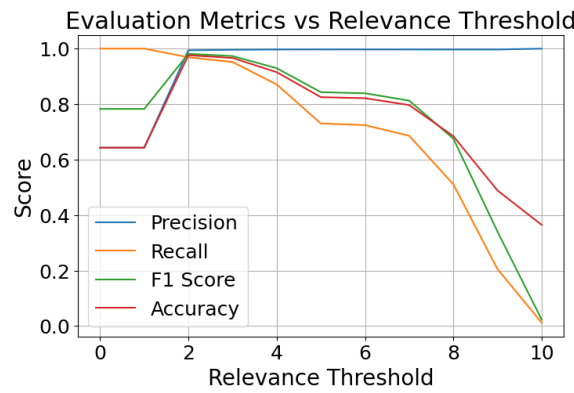
To evaluate the relevance of the citation context using LLMs, we utilized the same dataset as in the previous evaluation. It should be noted that the citation context used in this evaluation is from CiteScreener in order to make a comparison of the evaluation possible. The dataset consists of citation claims, the abstract of the cited paper, and the label stating if the context is "Related" or "Unrelated". The citation contexts are collected from open-access papers in the PMC PubMed library, utilizing a

specified number of characters before and after the citation in the paper. The dataset consists of 15,300 relevant and 8,008 irrelevant citations. To the best of our knowledge this is the only openly available dataset with citation claims; therefore, we use it to evaluate our methodology. The evaluation results are shown in Table 2.

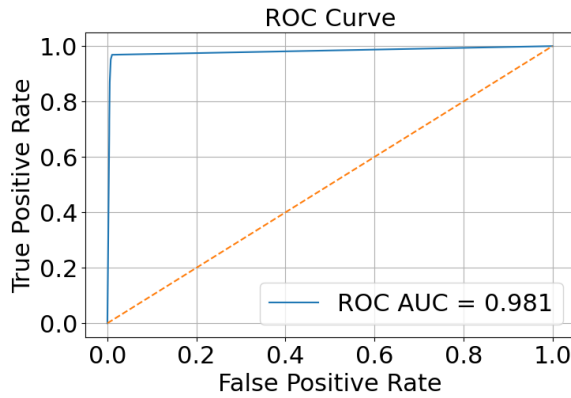
Table 2

Evaluation metrics for different LLMs

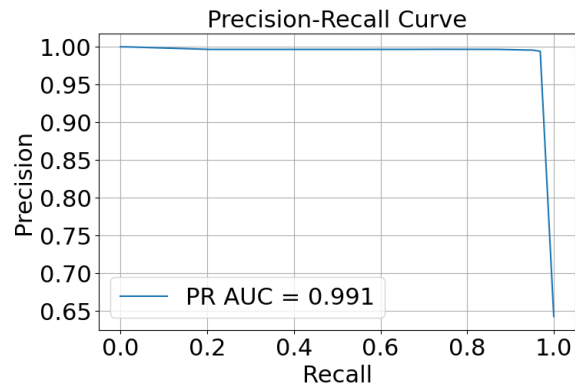
	Unrelated citations			Related citations			Accuracy
	Precision	Recall	F1	Precision	Recall	F1	
Qwen 2.5-7b	0.921	0.914	0.917	0.955	0.959	0.957	0.943
Gemma 2-9b	0.819	0.995	0.898	0.997	0.884	0.937	0.922
Gemma 2-27b	0.946	0.990	0.967	0.994	0.969	0.981	0.976
Mistral-7b	0.939	0.959	0.949	0.978	0.967	0.973	0.964
Llama 3.2-3b	0.886	0.634	0.739	0.833	0.957	0.891	0.846



(a) Metric variation across thresholds



(b) ROC curve.



(c) Precision-Recall curve.

Figure 2: Evaluation of thresholds for citation relevance detection.

We evaluate the approach across different thresholds of relevance score. The model used for evaluation is Gemma 2-27b, as it provides better results when compared to other models in terms of all evaluation metrics. The metrics for evaluation are shown in Fig. 2. Fig. 2a shows that increasing the threshold makes the classifier more conservative: precision remains high across most threshold values, while the other metrics reach their maximum at a threshold of 2 before declining thereafter.

The best result of the relevance analysis is achieved by the model Gemma 2-27b with an accuracy of 0.976 and an F1 score of 0.981 for relevant citations. This score is comparable to the best score from [12] which was achieved using the SBERT model, with accuracy and F1 of 0.985 and 0.981 respectively.

Fig. 2b shows the ROC curve, plotting the true positive rate against the false positive rate. The model achieves an *Area under the curve* (AUC) of 0.981, which demonstrates a strong capability of the model in distinguishing related and unrelated citations.

Fig. 2c shows the precision-recall curve, which is particularly informative for imbalanced datasets. The graph shows that the classifier maintains a high level of precision even at high recall values. These results suggest that a threshold of 2 is a good choice for separating relevant and irrelevant citations.

5. Misclassification analysis

Although the relevant citation analysis performs well on the evaluation dataset, with an F1 score of 0.981 for relevant citations, some citations are still misclassified. In this section, we analyze these misclassified cases to get a better understanding of the nature of the underlying errors, and categorize them into distinct types. The following sections describe the details of the categories and provide representative examples. Understanding these error types is essential for interpreting the results of the study, identifying the limitations of the proposed approach, and guiding future improvements. The misclassified citations are grouped into two categories as shown in the next section.

The original dataset comprised 23,308 rows, including 15,300 relevant and 8,008 irrelevant instances. However, the evaluation was performed on 22,381 rows (14,388 relevant and 7,993 irrelevant), as the LLM produced errors for 927 instances during the analysis, which were therefore excluded from the evaluation.

5.1. False positives - Unrelated citations classified as related

One type of error made by the model consists of false positives, i.e. unrelated citations classified as related. This category contains only 83 (approximately 1%) cases out of 7,993 evaluated citations. The number of false positives and false negatives is shown in the confusion matrix in Fig. 3, while representative examples of false positive cases are presented in Table 3.

In these cases the citation context is usually limited and does not contain keywords from any specific field. The citation is often attached to a brief, generic sentence. The LLM scores these cases as relevant because they contain generic language used in scientific studies, e.g. 'figure', 'studies', 'supplementary', 'reported'. For most of the cases, the relevance score is significantly lower as compared to true positives, shown in Table 4

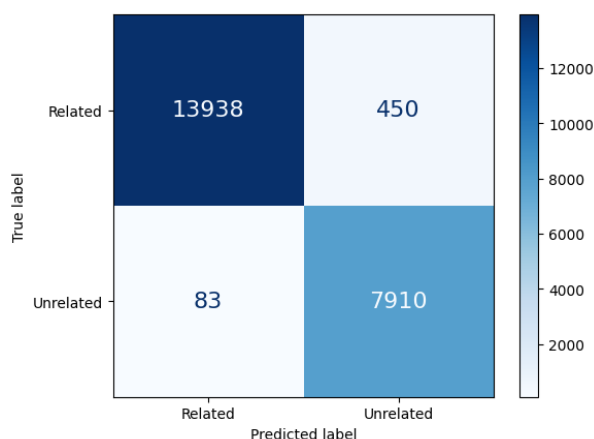


Figure 3: Confusion matrix for classification based on Gemma 2-27b model

These cases illustrate a limitation in the approach to computing abstract-level relevance between cited papers and citation contexts.

Table 3
Examples of false positives

Citation claim	Cited paper abstract	Predicted relevance score	Ground truth
This model is not what would be expected based on the claim of a recent study	Quantum computers may achieve speedups over their classical counterparts for solving linear algebra problems. However, in some cases—such as for low-rank ...	4	Irrelevant
For a few frames we also compared with FoxS	Abstract Linear regression analysis has become a fundamental tool in experimental sciences. We propose a new method for parameter estimation in linear models...	3	Irrelevant
This limitation might indicate that unlike the Sec complex	The aim of this paper is to develop a weighted functional linear Cox regression model that accounts for the association between a failure ...	4	Irrelevant

Table 4
Mean relevance scores of true and false positives

	Mean score(Scale: 1-10)
True Positives	6.84
False Positives	4.87

5.2. False negatives - Related citations classified as unrelated

This type of error by the LLM is more concerning than false positives, as it would flag genuine citations as irrelevant. It carries the risk of incorrectly accusing genuine scientists of misconduct. The model only classifies 450 (3.1%) out of 14,388 related citations as unrelated. Some examples of the false negatives can be seen in Table 5. The most common types of false negatives are the examples where a paper is cited, not to directly support the claim made in the citation context, but to provide a general context for the research area. These citations are legitimate, but are loosely connected to the specific claim in the citation context. As the LLM is instructed to determine the relevance between the claim and the abstract, it correctly deduces that the abstract does not directly support the narrow claim. This is a common practice in research papers, especially in the introduction section, as authors include citations to contextualize the academic foundation of their work where these citations do not necessarily explicitly confirm the claims made in the paper. The current approach has the limitation that it does not differentiate between direct and contextual citations.

Another kind of false negatives comes from the cited paper’s abstract. Some papers, especially older ones, have abstracts that are quite generic and do not convey specific findings of the study in the abstract. When the LLM evaluates the relevance of the citation, it generally provides it a lower score simply because the abstract is not detailed enough to verify the claim in the citing paper. As we are using abstracts instead of full text papers for citation relevance, this is a limitation of the approach.

5.3. Implications for deployment

As seen in the analysis of the misclassification results, it is clear that ReCitE should not be used as a standalone method for detecting integrity problems in scientific papers. It would be best used as a flagging tool, with a human in the loop, especially for cases where multiple citations in a paper are found to be irrelevant. A single irrelevant citation may just be a contextual reference or a parsing error, whereas a series of irrelevant references is a strong indication of integrity issues.

Secondly, the reliance of the approach on the accurate extraction of the citation context also greatly affects the results. Cases where the citation context is short and vague should be dealt with caution. Additionally, relying only on abstracts for relevance could sometimes be a disadvantage when the

Table 5

Examples of false negatives

Citation context	Cited paper abstract	Predicted relevance score	Ground truth
ELISA was performed in 384-well plates (Corning) as previously described	ABSTRACTV2/V3 conformational epitope antibodies that broadly neutralize HIV-1 (PG9 and PG16) have been recently described. Since an elicitation of previously known broadly...	1	Relevant
Livestock herding dogs display one of the most easily defined breed group-typical behaviors, characterized by an instinctive herding drive coupled with motor patterns that move herds in complex ways.	The navigation of commissural axons in the developing spinal cord has attracted multiple studies over the years. Many important concepts...	1	Relevant
These results indicate that S15 specifically promotes transposition by facilitating the assembly of the Cas12k-transposon recruitment complex, suggesting that it may function as an integral part of the type V CAST machinery. The 40S ribosomal protein S13 (RPS13) is the eukaryotic ortholog of the bacterial S15 protein	A system for naming ribosomal proteins is described that the authors intend to use in the future. They urge others to adopt it. The objective is to eliminate the confusion caused by the assignment of identical names to ribosomal proteins from different species that are unrelated in structure and function...	1	Relevant

abstract is not descriptive enough and is written in general terms. These limitations need to be addressed in future work.

6. Conclusion

This paper addresses citation anomalies in research papers and presents a framework for flagging irrelevant citations using LLMs. The study employs LLMs to detect and extract potentially irrelevant citations at the individual citation level. The citation context quality was evaluated manually on a random sample of 100 citations, with 51% having more accurate contexts and 36% having contexts similar to the baseline. Citation relevance was evaluated using multiple LLMs, with Gemma 2-27b producing the best results. The model achieved an F1 score of 0.967 for unrelated citations and 0.981 for related citations, with an ROC AUC of 0.981, demonstrating the model’s strong ability to discriminate between the two classes.

The scoring system of the pipeline uses a scale from 1 to 10, providing greater sensitivity than a binary classification approach. The study also provides a brief justification for each decision, explaining why a citation is considered relevant or irrelevant to the claim it is associated with. This enables human reviewers to better interpret and distinguish between contextual and irrelevant citations.

The misclassification analysis reveals that false positives mostly occur due to generic contextual citations and uninformative abstracts. Ideally, the approach should be used with a human-in-the-loop, as false negatives may result in legitimate citations being incorrectly flagged. The proposed approach, when combined with human oversight, could provide a useful tool for editors, reviewers, and research integrity sleuths to identify papers potentially associated with paper mill activity.

7. Limitations

Although our method can help flag integrity problems in suspicious research papers, there are still some limitations. The PDF parser used does not always perfectly label the references and paragraphs in a paper. As different journals/conferences have different formatting requirements, which is why, GROBID could not always perfectly extract and label the data, to provide good results for the irrelevant citation analysis.

The analysis of the irrelevant citations, produced some false positives while evaluating the approach. One should be extra cautious with these cases because flagging a genuine reference as being "irrelevant" could tarnish the reputation and damage careers of honest authors. Therefore, it is always advisable to have a human in the loop after the automated analysis to verify the results.

8. Future work

Future work would be to improve citation claim extraction by exploring alternate PDF parsing tools, hence, reducing sensitivity to GROBID's limitations. Another improvement could be retrieving the full text of the cited papers, in order to bypass uninformative abstracts. Additionally, creating a more detailed and extensive prompt could reduce the false negatives which arise from background references.

Additionally, the papers and their citations can be included in a knowledge graph which establishes relations between the entities. The papers and authors would be nodes, whereas authorship and citations would be the edges in the knowledge graph. The irrelevance of a citation would then be measured as graph distance or lack of a coherent path between citing and cited paper's nodes.

9. Data Availability

Supporting data is available from the authors upon reasonable request.

Acknowledgments

This study is a part of the FASCIFFT I 167 (Fake Science Journals and their Techniques) project at the Faculty of Informatics at Otto von Guericke University, Magdeburg, Germany. The project is funded by the state of Saxony Anhalt, Germany.

Declaration on Generative AI

During the preparation of this work, the authors used Claude and ChatGPT for grammar and spelling checks. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

- [1] B. A. Sabel, E. Knaack, G. Gigerenzer, et al., Fake publications in biomedical science: red-flagging method indicates mass production, *Naunyn-Schmiedeberg's Archives of Pharmacology* 399 (2026) 2943–2955. URL: <https://doi.org/10.1007/s00210-025-04275-9>. doi:10.1007/s00210-025-04275-9.
- [2] L. Parker, S. Boughton, L. Bero, J. A. Byrne, Paper mill challenges: past, present, and future, *Journal of Clinical Epidemiology* 176 (2024) 111549. URL: <https://www.sciencedirect.com/science/article/pii/S0895435624003056>. doi:<https://doi.org/10.1016/j.jclinepi.2024.111549>.
- [3] A. Abalkina, Publication and collaboration anomalies in academic papers originating from a paper mill: evidence from russia, *ArXiv abs/2112.13322* (2021). URL: <https://api.semanticscholar.org/CorpusID:245502369>.

- [4] The Center for Scientific Integrity, The retraction watch database, <http://retractiondatabase.org/>, 2018. Accessed: 2026-05-23.
- [5] Institute of Advanced Social Studies (IESA-CSIC), Retractable, <https://retractbase.csic.es/>, 2026. Open database of retracted and withdrawn scientific publications. Accessed: 2026-05-23.
- [6] A. Abalkina, R. Aquarius, B. Elisabeth, D. Bimler, D. Bishop, J. Byrne, G. Cabanac, A. Day, C. Labbé, N. Wise, ‘Stamp out paper mills’ — science sleuths on how to fight fake research, *Nature* 637 (2025) 1047–1050. URL: <https://www.nature.com/articles/d41586-025-00212-1>. doi:10.1038/d41586-025-00212-1.
- [7] G. Cabanac, C. Labbé, A. Magazinov, Tortured phrases: A dubious writing style emerging in science, arXiv preprint arXiv:2107.06751 (2021). URL: <https://arxiv.org/abs/2107.06751>.
- [8] G. Razis, K. Anagnostopoulos, O. Metaxas, S.-D. Stefanidis, H. Zhou, I. Anagnostopoulos, Papermill detection in scientific content, in: 2023 18th International Workshop on Semantic and Social Media Adaptation Personalization (SMAP), 2023, pp. 1–6. doi:10.1109/SMAP59435.2023.10255173.
- [9] M. Dadkhah, M. H. Oermann, M. Hegedüs, R. Raman, L. D. Dávid, Detection of fake papers in the era of artificial intelligence, *Diagnosis* 10 (2023) 390–397. URL: <https://doi.org/10.1515/dx-2023-0090>. doi:10.1515/dx-2023-0090.
- [10] A. K. Hussain, B. A. Sabel, M. Thiel, A. Nürnberger, Automated detection of fake biomedical papers: A machine learning perspective, in: International Conference on Enterprise Information Systems, 2025. URL: <https://api.semanticscholar.org/CorpusID:277684354>.
- [11] E. A. Fong, A. W. Wilhite, Authorship and citation manipulation in academic research, *PLOS ONE* 12 (2017) e0187394. URL: <https://doi.org/10.1371/journal.pone.0187394>. doi:10.1371/journal.pone.0187394.
- [12] Q. Liu, Y. Öztürk, T. Terkmani, F. Portet, C. Labbé, Citescreener: A pipeline for citation verification in digital libraries with datasets, in: 2025 ACM/IEEE Joint Conference on Digital Libraries (JCDL), 2025, pp. 158–166. doi:10.1109/JCDL67857.2025.00027.
- [13] J.-B. de la Broise, F. Sauerburger, E. Sayas, D.-M. Tecu, S. Meijere, M. Cuculovic, Cite lens: An ai tool for detecting out-of-scope and out-of-context citations, in: New Trends in Theory and Practice of Digital Libraries, volume 2694 of *Communications in Computer and Information Science*, Springer, Cham, 2026, pp. 25–34. URL: https://doi.org/10.1007/978-3-032-06136-2_3. doi:10.1007/978-3-032-06136-2_3.
- [14] Y. Huang, X. Yang, B. Zhang, X. Tang, Ce-kg: Citation enhanced knowledge graph via citation sentiment fusion and evidence tracing., in: ISWC (Industry/Doctoral Consortium/Posters/Demos), 2025, pp. 211–214.
- [15] H. A. Abid, M. A. Al-Obaidi, Irrelevant citations in medical and life science literature, *Baghdad Journal of Biochemistry and Applied Biological Sciences* 3 (2022) 1–3. URL: <https://doi.org/10.47419/bjbabs.v3i01.120>. doi:10.47419/bjbabs.v3i01.120.
- [16] Grobid, <https://github.com/grobidOrg/grobid>, 2008.

Appendices

A. Prompt to the LLMs

The prompt provided to the LLM is detailed as the it has multiple tasks to perform. The prompt is divided into six sections:

1. Assigning the LLM a role

ROLE: You are an objective Research Analyst. And your job is to isolate the text relating to the specific claim being made in the CONTEXT PARAGRAPH and then evaluate how well the ABSTRACT of the cited reference supports that claim.

2. Defining the task

TASK:

Evaluate the strength of alignment between a specific claim made in a research paper and the abstract of the cited reference.

3. Providing required data

DATA:

CONTEXT PARAGRAPH: context

CITATION TO FOCUS ON: ref_id_to_find

REFERENCE TITLE: title

REFERENCE ABSTRACT: abstract

4. Providing instructions

INSTRUCTIONS:

1. Identify the specific assertion in the CONTEXT PARAGRAPH that is directly linked to the citation ref_id_to_find. Be mindful that the paragraph may contain multiple sentences and only a subset of them may be relevant to the citation. Focus on isolating the claim that is most likely being supported by the cited reference. Also be mindful that if the context contains multiple citations, the claim you identify should be just the sentence that contains the specific citation ref_id_to_find.

2. Rate the strength of the relationship between that assertion and the findings in the ABSTRACT on a scale of 1 to 10.

5. Defining scoring Rubric

SCORING RUBRIC:

10: The abstract directly confirms the claim with specific data or conclusions.

7-9: The abstract provides strong supporting evidence or thematic alignment.

4-6: The abstract is generally related but does not explicitly confirm the specific assertion.

1-3: The abstract is on a different topic or contradicts the claim.

OUTPUT:

Return ONLY a JSON object with this structure:

```
{
  "identified_claim": "The specific sentence/claim being supported",
  "alignment_score": int,
  "brief_justification": "1-sentence explanation of the score"
}
```

B. Examples of citation context extraction comparison between ReCitE and CiteScreener

Table B1: A case where ReCitE is more accurate

Category	Citation context from our approach	Citation context from CiteScreener	Abstract of cited paper	Paragraph in main paper where citation is present (relevant citation in bounding box)	Citation number
Our approach provides a more accurate context	The term super-enhancer (SE) has been coined to describe clusters of elements bearing the bioinformatic signature of enhancers, which are enriched for particularly high levels of H3K27Ac, TF recruitment, and recruitment of transcriptional co-activators such as the Mediator complex.	Generally, enhancers are classified bioinformatically based on chromatin accessibility, TF occupancy, and enrichment for particular histone modifications including H3K4Me1 and H3K27Ac. The term super-enhancer (SE) has been coined to describe clusters of elements bearing the bioinformatic signature of enhancers, which are enriched for particularly high levels of H3K27Ac, TF recruitment, and recruitment of transcriptional co-activators such as the Mediator complex.	Master transcription factors Oct4, Sox2, and Nanog bind enhancer elements and recruit Mediator to activate much of the gene expression program of pluripotent embryonic stem cells (ESCs). We report here that the ESC master transcription factors form unusual enhancer domains at most genes that control the pluripotent state. These domains, which we call super-enhancers, consist of clusters of enhancers that are densely occupied by the master regulators and Mediator. Super-enhancers differ from typical enhancers in size, transcription factor density and content, ability to activate transcription, and sensitivity to perturbation. Reduced levels of Oct4 or Mediator cause preferential loss of expression of super-enhancer-associated genes relative to other genes, suggesting how changes in gene expression programs might be accomplished during development. In other more differentiated cells, super-enhancers containing cell-type-specific master transcription factors are also found at genes that define cell identity. Super-enhancers thus play key roles in the control of mammalian cell identity.	Enhancers are regions of DNA that recruit transcription factors (TFs) and activate expression of target genes in a celltype-specific manner. ^{1–3} Generally, enhancers are classified bioinformatically based on chromatin accessibility, TF occupancy, and enrichment for particular histone modifications including H3K4Me1 and H3K27Ac. The term super-enhancer (SE) has been coined to describe clusters of elements bearing the bioinformatic signature of enhancers, which are enriched for particularly high levels of H3K27Ac, TF recruitment, and recruitment of transcriptional co-activators such as the Mediator complex. ⁴ SEs are often regulators of cell identity genes and are frequently mutated in association with complex traits and genetic diseases. ^{5–9}	4

Table B2: A case where both contexts are similar and accurate

Category	Citation context from our approach	Citation context from CiteScreener	Abstract of cited paper	Paragraph in main paper where citation is present (relevant citation in bounding box)	Citation number
context almost identical and accurate	Enhancers are regions of DNA that recruit transcription factors (TFs) and activate expression of target genes in a celltype-specific manner. [1][2][3]	Enhancers are regions of DNA that recruit transcription factors (TFs) and activate expression of target genes in a cell-type-specific manner.	A class of cis-regulatory elements, called enhancers, play a central role in orchestrating spatiotemporally precise gene-expression programs during development. Consequently, divergence in enhancer sequence and activity is thought to be an important mediator of inter- and intra-species phenotypic variation. Here, we give an overview of emerging principles of enhancer function, current models of enhancer architecture, genomic substrates from which enhancers emerge during evolution, and the influence of three-dimensional genome organization on long-range gene regulation. We discuss intricate relationships between distinct elements within complex regulatory landscapes and consider their potential impact on specificity and robustness of transcriptional regulation.	Enhancers are regions of DNA that recruit transcription factors (TFs) and activate expression of target genes in a cell-type-specific manner. [1–3] Generally, enhancers are classified bioinformatically based on chromatin accessibility, TF occupancy, and enrichment for particular histone modifications including H3K4Me1 and H3K27Ac. The term super-enhancer (SE) has been coined to describe clusters of elements bearing the bioinformatic signature of enhancers, which are enriched for particularly high levels of H3K27Ac, TF recruitment, and recruitment of transcriptional co-activators such as the Mediator complex. ⁴ SEs are often regulators of cell identity genes and are frequently mutated in association with complex traits and genetic diseases. ^{5–9}	1

Table B3: A case where ReCite is less accurate

Category	Citation context from our approach	Citation context from CiteScreeener	Abstract of cited paper	Paragraph in main paper where citation is present (relevant citation in bounding box)	Citation number
our method is less precise	We analysed time-to-event outcomes using the Kaplan-Meier product limit estimator, log-rank tests for differences between groups, and Cox-proportional hazards models; these were displayed with the KMunicate format. 11 We tested assumptions of proportional hazards using Schoenfeld residuals.	We analysed time-to-event outcomes using the Kaplan-Meier product limit estimator, log-rank tests for differences between groups, and Cox-proportional hazards models; these were displayed with the KMunicate format.	ObjectivesTo examine reactions to the proposed improvements to standard Kaplan-Meier plots, the standard way to present time-to-event data, and to understand which (if any) facilitated better depiction of (1) the state of patients over time, and (2) uncertainty over time in the estimates of survival.DesignA survey of stakeholders' opinions on the proposals.SettingA web-based survey, open to international participation, for those with an interest in visualisation of time-to-event data.Participants174 people participated in the survey over a 6-week period. Participation was global (although primarily Europe and North America) and represented a wide range of researchers (primarily statisticians and clinicians).Main outcome measuresTwo outcome measures were of principal importance: (1) participants' opinions of each proposal compared with a standard Kaplan-Meier plot; and (2) participants' overall ranking of the proposals (including the standard).ResultsMost proposals were more popular than the standard Kaplan-Meier plot. The most popular proposals in the two categories, respectively, were an extended table beneath the plot depicting the numbers at risk, censored and having experienced an event at periodic timepoints, and CIs around each Kaplan-Meier curve.ConclusionsThis study produced a high response number, reflecting the importance of graphics for time-to-event data. Those producing and publishing Kaplan-Meier plots both authors and journals should, as a starting point, consider using the combination of the two favoured proposals.	In the primary efficacy analysis, we calculated the absolute between-group difference (with 95% CI) in the proportion of participants with a favourable outcome, adjusted for HIV status and randomisation protocol, using Cochran-Mantel-Haenszel weights.10 Non-inferiority was shown if the upper boundary of the 95% CI was less than 10% in both the modified intention-to-treat (mITT) and per-protocol populations. The mITT population included all randomly assigned participants with a positive culture for M tuberculosis at screening or randomisation, except for participants with isolates taken before randomisation who were subsequently found to be susceptible to rifampicin or resistant to both fluoroquinolones and second-line injectables on phenotypic drug-susceptibility testing. The per-protocol population is the same as the mITT population with the exclusion of participants who did not complete a protocol-adherent course of treatment, other than for treatment failure, change of treatment for an adverse event, or death (appendix pp 16–17). We did one-sided tests for non-inferiority and calculations of the 95% CIs using the Wald standard error. Prespecified tests for superiority were done when non-inferiority was shown. The proportion of participants meeting each safety outcome was calculated similarly in the safety population, which comprised all participants who received at least one dose of a trial medication, with two-sided tests of superiority. We included a Bayesian interpretation of the primary outcome as a supplementary analysis. Prespecified sensitivity analyses of the primary efficacy analysis included analyses that were unadjusted, adjusted for important baseline characteristics, and where the definition of an unfavourable outcome was modified (appendix p 25). The primary efficacy analysis was also repeated in subgroups according to prespecified baseline characteristics (appendix p 26). We analysed time-to-event outcomes using the Kaplan-Meier product limit estimator, log-rank tests for differences between groups, and Cox-proportional hazards models; these were displayed with the KMunicate format.¹¹ We tested assumptions of proportional hazards using Schoenfeld residuals.	11

Table B4: A case where ReCitE is not correct

Category	Citation context from our approach	Citation context from CiteScreener	Abstract of cited paper	Paragraph in main paper where citation is present (relevant citation in bounding box)	Citation number
Our method is not correct	The term super-enhancer (SE) has been coined to describe clusters of elements bearing the bioinformatic signature of enhancers, which are enriched for particularly high levels of H3K27Ac, TF recruitment, and recruitment of transcriptional co-activators such as the Mediator complex.	Despite extensive analysis, it remains unclear whether SEs are merely groups of independent classical enhancers or whether they are cooperatives containing functionally distinct element types.	AbstractGene expression is precisely controlled in a stage and cell-type-specific manner, largely through the interaction between cis-regulatory elements and their associated trans-acting factors. Where these components aggregate in promoters and enhancers, they are able to cooperate to modulate chromatin structure and support the engagement in long-range 3D superstructures that shape the dynamics of a cell's genomic architecture. Recently, the term "super-enhancer"™ has been introduced to describe a hyper-active regulatory domain comprising a complex array of sequence elements that work together to control the key gene networks involved in cell identity. Here, we survey the unique characteristics of super-enhancers compared to other enhancer types and summarize the recent advances in our understanding of their biological role in gene regulation. In particular, we discuss their capacity to attract the formation of phase-separated condensates, and capacity to generate three-dimensional genome structures that precisely activate their target genes. We also propose a multi-stage transition model to explain the evolutionary pressure driving the development of super-enhancers in complex organisms, and highlight the potential for involvement in tumorigenesis. Finally, we discuss more broadly the role of super-enhancers in human health disorders and related potential in therapeutic interventions.	Despite extensive analysis, it remains unclear whether SEs are merely groups of independent classical enhancers or whether they are cooperatives containing functionally distinct element types.^[10–14] SEs may not all be mechanistically the same. This has become a highly debatable topic as the genetic dissection of SEs is challenging, given that many studied loci regulate genes that control complex transcriptional and epigenetic programs. Disruption of such pathways makes it difficult to separate changes in SE-regulated gene expression from associated changes in cell lineage and differentiation. Most previous studies analyzing SEs have drawn conclusions from deleting just one or two constituent elements, and many studies have relied on artificial reporter-based assays divorced from their functionally relevant chromatin contexts.¹⁵ To date, a number of studies have dissected SEs^{16–22} with conflicting conclusions. More rigorous analysis is essential to determine their true nature.	14