Detecting Reference Errors in Scientific Literature with Large Language Models

Anonymous ACL submission

Abstract

Reference errors, such as citation and quotation errors, are common in scientific papers. Such errors can result in the propagation of inaccurate information, but are difficult and time-consuming to detect, posing a significant challenge to scientific publishing. To support automatic detection of reference errors, this work evaluated the ability of large language models in OpenAI's GPT family to detect quotation errors. Specifically, we prepared an expert-annotated, general-domain dataset of statement-reference pairs from journal articles. Large language models were evaluated in different settings with varying amounts of reference information provided by retrieval augmentation. Our results showed that large language models are able to detect erroneous citations with limited context and without fine-tuning. This study contributes to the growing literature that seeks to utilize artificial intelligence to assist in the writing, reviewing, and publishing of scientific papers. Potential avenues for further improvements in this task are also discussed.

1 Background and Introduction

Researchers cite literature as references and supporting evidence when reporting their work in papers. The reliability of referencing is usually taken for granted. However, previous citation verification studies in multiple scientific domains have revealed that reference errors of varying degrees are common in scientific papers, with prevalence rates ranging from 11% to 41%, depending on the domain, journal, and methodology (Goldberg et al., 1993; Lee and Lee, 1999; Fenton et al., 2000; Gosling et al., 2004; Lukić et al., 2004; Todd et al., 2010; Jergas and Baethge, 2015; Mogull, 2017; Armstrong et al., 2018; Smith and Cumberledge, 2020; Pavlovic et al., 2021; Cobb et al., 2024). Reference errors could result in the propagation of inaccurate information (Smith and Cumberledge, 2020), undermining the credibility of scientific research and sometimes leading to serious consequences (Pavlovic et al., 2021). For example, hundreds of uncritical citations of a 1980 letter published in the New England Journal of Medicine may have contributed to the opioid crisis in the United States (Leung et al., 2017).

Previous studies defined two major types of reference errors: citation errors and quotation errors. Citation errors usually refer to typographical errors in referencing, such as incorrect reference information (e.g., incorrect authors, title, journal, or year) or the erroneous arrangement of references (Smith and Cumberledge, 2020). Citation errors have become less common in the era of digitization and citation managers, although these same factors may enable the propagation of pre-existing citation errors. In contrast, a quotation error specifically refers to the situation where a reference fails to support the statement for which it is cited (Smith and Cumberledge, 2020). Notably, these two definitions are not mutually exclusive, as an incorrect assignment of reference indices can manifest as a quotation error in an individual statement-reference pair. Recent studies on referencing errors sometimes use different terms for quotation errors, such as "content errors" (Mogull, 2017), "inaccurate citations" (Pavlovic et al., 2021), or "miscitations" (Cobb et al., 2024).

Quotation errors are difficult and time-consuming to detect, and they often require domain expertise when comparing a statement to relevant information in the reference article (Smith and Cumberledge, 2020). Previous studies on reference errors typically utilized domain experts to manually examine samples of scientific papers. The difficulty of detecting quotation errors poses a significant challenge to scientific publishing, as it requires additional efforts by editors and reviewers in peer review. Given the exploding number of papers being published each year, this task is becoming increasingly demanding.

Fully substantiated: The reference article fully substantiates the relevant part of the statement.
Partially substantiated: According to the reference article, there is a minor error in the statement, but the error does not invalidate the purpose of the statement.
Unsubstantiated: The reference part does not substantiate any part of the statement. This could be because the statement is contradictory to, unrelated to, or simply missing from the reference article.

Table 1: Label definitions.

Label	**n (%)**
Unsubstantiated	112 (44.8)
Partially substantiated	14 (5.6)
Fully substantiated	124 (49.6)
Domain	**n (%)**
Biology or Medicine	85 (34.0)
Chemistry or Material Science	57 (22.8)
Physics	26 (10.4)
Social Science	26 (10.4)
Earth or Environmental Science	24 (9.6)
Engineering	17 (6.8)
Computer Science	15 (6.0)
Reference Availability	**n (%)**
Has abstract	242 (96.8)
Has PDF	244 (97.6)
Has abstract or PDF	250 (100)

Table 2: Dataset characteristics.

Recent advances in natural language processing (NLP) have demonstrated astounding capabilities of large language models (LLMs) to perform various types of text-based tasks (Ouyang et al., 2022; Touvron et al., 2023; OpenAI et al., 2024; Google, 2024), providing a strong baseline for application in the real world. Researchers have also started exploring ways NLP can assist with paper writing and peer review (Kuznetsov et al., 2024). However, none of these studies examined reference error detection. To fill this gap and encourage future attempts to automate reference error detection, this study performed a general-domain evaluation of the capability of LLMs to detect quotation errors in scientific papers.

2 Task and Related Work

The quotation error detection task of this work is defined as follows: given a statement s and a reference article r that the statement cites, a model should predict a label $f(s, r) \in$ {Fully substantiated, Partially substantiated, Unsubstantiated} to indicate whether the statement-reference pair contains a quotation error. The names and definitions of the labels follow previous citation verification studies (Smith and Cumberledge, 2020; Cobb et al., 2024). Complete definitions of the labels are listed in Table 1.

There have been some studies in the computer science domain that both utilized citations and employed a similar task or methodology (to ours) but were for different scenarios. Motivated by fact-checking claims related to COVID-19, Wadden et al. (2020) proposed a scientific claim verification task in which each atomic statement/claim is verified against a corpus of paper abstracts to determine whether it is supported or refuted by the literature. Several teams of researchers, including Wadden et al. themselves, subsequently developed models for this shared task using the same dataset (Pradeep et al., 2021; Li et al., 2021; Zhang et al., 2021; Wadden et al., 2022). Notably, the scientific claim verification task labels a claim that is neither supported nor refuted by the corpus as "Not Enough Information". Furthermore, the scope of the scientific claim verification task was limited to biomedicine. In terms of LLM-generated citations, Wu et al. examined the quality of web-based, instead of peer-reviewed article based, citations generated by LLMs in response to medical questions (Wu et al., 2024).

3 Dataset and Experiments

The evaluation dataset is available at [url]. Distributions of labels, domains, and reference availability in the dataset are summarized in Table 2. Some examples of quotation errors in the dataset as well as data collection and quality control details are provided in Appendix A and B, respectively.

LLMs were given task instructions and information about both the citing article and the reference. The prompt template (Appendix C) was finalized before the start of the experiment. To investigate the impact of information completeness, LLMs were tested in 3 settings with different amounts of information from the reference: (1) With only the title of the reference provided; (2) With both the title and abstract provided; (3) With the title, abstract, and excerpts provided.

Local retrieval of excerpts from the main body of a reference followed a 3-step retrieval-augmented generation (RAG) (Gao et al., 2024) pipeline. First, the full text of a reference was extracted from its PDF file by GROBID[1]. The extracted full text was then split into 256-token chunks with 20-token overlaps using LlamaIndex[2]. In the experiment, the embeddings of the chunks were compared to that of the input statement using LlamaIndex, and the top 3 chunks with the highest match to the statement were retrieved and included in the prompt to the LLMs.

Additionally, we included OpenAI's Assistant API in the experiment since it provides a proprietary RAG workflow for LLMs to utilize information in PDF attachments. The assistant received the same prompt format containing the title of the reference, plus instructions to see the attached PDF file for the full text of the reference.

Three LLMs in OpenAI's GPT family were evaluated in the experiment: `gpt-3.5-turbo-0125`, `gpt-4-0125-preview`, and `gpt-4o-2024-05-13`. LLMs were prompted to respond with a JSON object containing a predicted label and an explanation for their selection. All LLM experiments were conducted using OpenAI's Python API with temperature set to 0. Model performance was measured by label accuracy. Since Wadden et al.'s (2020) scientific claim verification task also applies to single statement-reference pairs, relevant models were also tested on our dataset for comparison.

4 Results

Table 3 provides by-class and overall performance of the models in different settings. GPT-4 Turbo and GPT-4o performed much better than GPT-3.5 Turbo at detecting Unsubstantiated cases, especially when information about the reference article is limited. Provision of more information or the Assistant pipeline did not necessarily improve LLM performance, especially for GPT-3.5 Turbo which performed much better on Fully substantiated cases in the title-only setting than with more information. The models for Wadden et al.'s (2020) scientific claim verification task predicted all statement-reference pairs in our dataset as "Not Enough Information" when using abstracts as references, suggesting that these models may not be directly applicable to our quotation error detection task. Therefore, they are excluded from Table 3.

[1] https://github.com/kermitt2/grobid

[2] https://github.com/jerryjliu/llama_index

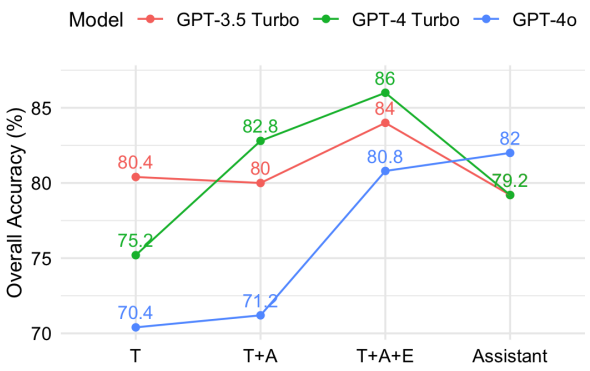

Figure 1: Two-class performance of LLMs. "T", "A", "E" stand for title, abstract, and excerpts, respectively.

Error analysis revealed that LLMs with low accuracy on Fully substantiated pairs predicted most of them as Partially substantiated. This was frequently the case when a statement contained multiple sub-claims, and LLMs were "confused" by the ambiguity in the span of the part to be verified. For instance, the statement in error example 1 of Appendix D cites the reference article to support its data from South China only. GPT-4 Turbo assumed that the reference should substantiate the entire statement and predicted this case as Partially substantiated because the reference does not support other locations mentioned (Greenland and Svalbard).

Considering both the rareness of Partially substantiated pairs in the dataset and their relatively low importance from a practical perspective, we then merged Partially and Fully substantiated predictions and performed a secondary analysis of model performance. Two-class performance (Figure 1) clearly shows that more reference information improved LLM predictions in general.

A scrutiny of errors and explanations of LLMs revealed distinct model behaviors that are consistent with the trends in model performance. LLM explanations typically begin with a summary of the two articles followed by a comparison of them, clearly showing how a model made its prediction. When only the title of the reference is given, LLMs are forced to rely on superficial relations between the statement and the reference. GPT-3.5 Turbo tends to be permissive as long as the two articles seem to belong to the same research area. When more information is provided, GPT-3.5 Turbo would still rely on such superficial relations,

Model	Given Information	Accuracy by Class			Accuracy
		Un	Partially	Fully	Overall
GPT-3.5 Turbo	Title	64.3	14.3	73.4	66.0
	Title + Abstract	84.8	57.1	30.6	56.4
	Title + Abstract + Excerpts	79.5	57.1	30.6	54.0
	Title + PDF (Assistant)	79.5	14.3	63.7	68.0
GPT-4 Turbo	Title	89.3	14.3	36.3	58.8
	Title + Abstract	89.3	35.7	39.5	61.6
	Title + Abstract + Excerpts	83.9	21.4	62.9	70.0
	Title + PDF (Assistant)	84.8	35.7	55.6	67.6
GPT-4o	Title	90.2	14.3	29.8	56.0
	Title + Abstract	92.0	28.6	13.7	49.6
	Title + Abstract + Excerpts	86.6	50.0	34.7	58.8
	Title + PDF (Assistant)	83.9	21.4	58.9	68.0

Table 3: LLM performance on quotation error detection.

and additional text from the reference that is not directly related to the statement would cause GPT-3.5 Turbo to regard the two articles as not so related. This also explains why GPT-3.5 Turbo's performance on Unsubstantiated cases increased dramatically when more information was provided. Unlike GPT-3.5 Turbo, GPT-4 Turbo and GPT-4o tend to be much stricter in error detection, paying attention to small differences between the statement and the reference, although such differences are usually acceptable for humans since researchers frequently rephrase, abstract, and generalize statements from references, or even combine them with additional information. For instance, the statement in error example 2 of Appendix D mentions copper as an example of transition metal catalysts while introducing Grignard reagents, and GPT-4 Turbo predicted this statement as Partially substantiated because the reference article about Grignard reagents does not mention copper explicitly.

Considering that hallucination can be a major threat to LLM generation, we carefully reviewed LLM-generated explanations during error analysis. No significant sign of hallucination was found.

5 Discussion

Scientific scholarship depends heavily on the ability to trust, verify, replicate, or refute prior research. The increasing availability of electronic publications, advanced search, and citation managers has expanded the ability of researchers to cite existing findings, yet this process is still prone to the introduction of reference errors.

This study contributes to the growing literature that seeks to utilize artificial intelligence to assist in the writing, reviewing, and publishing of scientific papers by providing a general-domain dataset of statement-reference pairs to encourage future studies in this field. We found that LLMs can detect quotation errors in scientific papers with limited context and without fine-tuning. We also quantified the relative contributions of model versions and increasing levels of context which could affect cost and speed in a production environment. During data collection, we also observed that a higher frequency of major reference errors in a paper appears to reflect citation manipulation behaviors. Given the capability of LLMs to detect quotation errors, this work may also contribute to reducing academic misconduct and fake papers that are polluting the scientific literature (Sanderson, 2024).

Our results reveal misalignment in the comprehension of reference errors between humans and LLMs, which points the way to several promising avenues for further research. Interrogation into the features and mechanisms of quotation error detection may yield insights into which models, training corpora, or fine-tuning schemes would further improve performance. Prompt engineering (e.g. few-shot learning, multi-step chain-of-thought reasoning), ensemble methods, and other approaches are likely to demonstrate further gains in accuracy. Certain domains or types of complex statements appear more resistant to accurate classification; identifying these would help improve performance, characterize system limitations, and develop products to aid comprehensive and efficient systems for editorial and fact-checking tasks.

6 Limitation

Our study was limited in size and scope. During the progress of the study, new variants of Claude, Gemini, Llama, and GPT became available, some of which support a long enough context window to accept an entire reference article as input. Other models such as AI21 have more extensive token dictionaries. It is likely that the out-of-the-box performance on this task will continue to advance.

Other limitations included our reliance on publicized and crowd-sourced datasets, the use of a simple sentence pair annotation scheme, and the treatment of all reference pairs as being equivalent despite potential multiple rationales for citation. In addition, all statement-references pairs in our dataset came from journal articles and were predominantly in natural science domains. Such characteristics limit the generalizability of our findings to papers that are published through other channels (e.g., conferences and preprint platforms) or in research domains that are underrepresented in this study. Further study would benefit from an expanded dataset labeled by multiple domain experts, and a randomly sampled and multilingual dataset indexed by year.

Acknowledgments

The authors received no funding for this study.

References

Michael F. Armstrong, Joseph H. Conduff III, John E. Fenton, and Daniel H. Coelho. 2018. Reference errors in otolaryngology–head and neck surgery literature. *Otolaryngology–Head and Neck Surgery*, 159(2):249–253.

John C. Buchan, John Norris, and Hannah Kuper. 2005. Accuracy of referencing in the ophthalmic literature. *American Journal of Ophthalmology*, 140(6):1146–1148.

Cory L. Cobb, Brianna Crumly, Pablo Montero-Zamora, Seth J. Schwartz, and Charles R. Martínez Jr. 2024. The problem of miscitation in psychological science: Righting the ship. *American Psychologist*, 79(2):299–311.

J. E. Fenton, H. Brazier, A. De Souza, J. P. Hughes, and D. P. McShane. 2000. The accuracy of citation and quotation in otolaryngology/head and neck surgery journals. *Clinical Otolaryngology and Allied Sciences*, 25(1):40–44.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Computing Research Repository*, arXiv:2312.10997.

Richard Goldberg, Edward Newton, Julie Cameron, Raymond Jacobson, Linda Chan, W. Richard Bukata, and Amine Rakab. 1993. Reference accuracy in the emergency medicine literature. *Annals of Emergency Medicine*, 22(9):1450–1454.

Gemini Team Google. 2024. Gemini: A family of highly capable multimodal models. *Computing Research Repository*, arXiv:2312.11805.

Cameron McR Gosling, Melainie Cameron, and Peter F. Gibbons. 2004. Referencing and quotation accuracy in four manual therapy journals. *Manual Therapy*, 9(1):36–40.

Helen H. G. Handoll and Greg Atkinson. 2015. Snowballing citations. *BMJ (Clinical research ed.)*, 351:h6309.

Hannah Jergas and Christopher Baethge. 2015. Quotation accuracy in medical journal articles—a systematic review and meta-analysis. *PeerJ*, 3:e1364.

Ilia Kuznetsov, Osama Mohammed Afzal, Koen Dercksen, Nils Dycke, Alexander Goldberg, Tom Hope, Dirk Hovy, Jonathan K. Kummerfeld, Anne Lauscher, Kevin Leyton-Brown, Sheng Lu, Mausam, Margot Mieskes, Aurélie Névéol, Danish Pruthi, Lizhen Qu, Roy Schwartz, Noah A. Smith, Thamar Solorio, Jingyan Wang, Xiaodan Zhu, Anna Rogers, Nihar B. Shah, and Iryna Gurevych. 2024. What can natural language processing do for peer review? *Computing Research Repository*, arXiv:2405.06563.

Sung Yul Lee and Jong Suk Lee. 1999. A survey of reference accuracy in two asian dermatologic journals (the journal of dermatology and the korean journal of dermatology). *International Journal of Dermatology*, 38(5):357–360.

Pamela T. M. Leung, Erin M. Macdonald, Matthew B. Stanbrook, Irfan A. Dhalla, and David N. Juurlink. 2017. A 1980 letter on the risk of opioid addiction. *The New England Journal of Medicine*, 376(22):2194–2195.

Xiangci Li, Gully Burns, and Nanyun Peng. 2021. A paragraph-level multi-task learning model for scientific fact-verification. *Computing Research Repository*, arXiv:2012.14500.

Ivan Krešimir Lukić, Anita Lukić, Vicko Glunčić, Vedran Katavić, Vladimira Vučenik, and Ana Marušić. 2004. Citation and quotation accuracy in three anatomy journals. *Clinical Anatomy*, 17(7):534–539.

Scott A. Mogull. 2017. Accuracy of cited "facts" in medical research articles: A review of study methodology and recalculation of quotation error rate. *PLOS ONE*, 12(9):e0184727.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, . . . , and Barret Zoph. 2024. Gpt-4 technical report. *Computing Research Repository*, arXiv:2303.08774.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Computing Research Repository*.

Vedrana Pavlovic, Tracey Weissgerber, Dejana Stanisavljevic, Tatjana Pekmezovic, Ognjen Milicevic, Jelena Milin Lazovic, Andja Cirkovic, Marko Savic, Nina Rajovic, Pavle Piperac, Nemanja Djuric, Petar Madzarevic, Ana Dimitrijevic, Simona Randjelovic, Emilija Nestorovic, Remi Akinyombo, Andrija Pavlovic, Ranine Ghamrawi, Vesna Garovic, and Natasa Milic. 2021. How accurate are citations of frequently cited papers in biomedical literature? *Clinical science (London, England: 1979)*, 135(5):671–681.

Ronak Pradeep, Xueguang Ma, Rodrigo Nogueira, and Jimmy Lin. 2021. Scientific claim verification with vert5erini. In *Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis*, pages 94–103. Association for Computational Linguistics.

Katharine Sanderson. 2024. Science's fake-paper problem: high-profile effort will tackle paper mills. *Nature*, 626(7997):17–18.

Neal Smith and Aaron Cumberledge. 2020. Quotation errors in general science journals. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 476(2242):20200538.

Peter A. Todd, James R. Guest, Junxiu Lu, and Loke Ming Chou. 2010. One in four citations in marine biology papers is inappropriate. *Marine Ecology Progress Series*, 408:299–303.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Computing Research Repository*, arXiv:2302.13971.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550. Association for Computational Linguistics.

David Wadden, Kyle Lo, Lucy Lu Wang, Arman Cohan, Iz Beltagy, and Hannaneh Hajishirzi. 2022. Multivers: Improving scientific claim verification with weak supervision and full-document context. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 61–76. Association for Computational Linguistics.

Kevin Wu, Eric Wu, Ally Cassasola, Angela Zhang, Kevin Wei, Teresa Nguyen, Sith Riantawan, Patricia Shi Riantawan, Daniel E. Ho, and James Zou. 2024. How well do llms cite relevant medical references? an evaluation framework and analyses. *Computing Research Repository*, arXiv:2402.02008.

Zhiwei Zhang, Jiyi Li, Fumiyo Fukumoto, and Yanming Ye. 2021. Abstract, rationale, stance: A joint model for scientific claim verification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3580–3586. Association for Computational Linguistics.

A Examples of Quotation Errors

Example 1: Partially substantiated
Reason: Incorrect number
Statement:
The relative topography of the origins of the renal aa. from the aorta was found to be variable, thus confirming earlier observations from angiographic studies in which the right a. arose more proximally in 65% of cases [citation mark].
Reference:
The ostium of the right renal a. was more cranial than the ostium of the left renal a. (53.3%) ...
Label source: (Smith and Cumberledge, 2020)

Example 2: Partially substantiated
Reason: Missing condition
Statement:
At very high crystallinities (very low melt permeabilities), gas-generated overpressures can fracture and brecciate the solidifying mush [citation mark].
Reference:
... for overpressure to develop, the combined crystallization expansion of plagioclase and exsolution of water is required to overcome the crystallization contraction of the other crystallizing phases... The requirement that plagioclase joins the liquidus for crystallization overpressure to develop is significant...
Label source: (Smith and Cumberledge, 2020)

Example 3: Unsubstantiated
Reason: Irrelevant reference
Statement:

There is an extremely significant positive correlation between students' behavioral participation, cognitive participation, emotional participation, and their variables and translation performance in high school translation classrooms [citation mark].
Reference:
(Title) Strong negative correlation between codon usage bias and protein structural disorder impedes protein expression after codon optimization
Label source: PubPeer

Example 4: Unsubstantiated
Reason: Contradiction
Statement:
We know that oxytocin is released when listening to music, and importantly, oxytocin is increased when participating in several forms of group singing, including improvisation [citation mark].
Reference:
We quantified mood and salivary OXT and cortisol (CORT) concentrations... Happiness was increased, and worry and sadness as well as salivary CORT concentrations were reduced, after both choir and solo singing.
Label source: Journal correction

B Data Collection Details

Statement-reference pairs in the dataset were collected through the following channels: (1) 163 (65.2%) pairs are from previous citation verification studies that either provided traceable examples of quotation errors or shared annotated datasets (Lee and Lee, 1999; Fenton et al., 2000; Gosling et al., 2004; Lukić et al., 2004; Buchan et al., 2005; Handoll and Atkinson, 2015; Smith and Cumberledge, 2020). Since the annotated datasets may not provide sufficient details about the specific sentence to be verified, we reviewed each citing article to locate the statement to which the label was assigned. (2) 80 (32.0%) pairs are from PubPeer[3], a platform for researchers to leave comments on others' publications. While collecting annotations from this source, we cross-referenced papers retracted in 2022 and 2023 due to "concerns or issues about referencing or attributions" in the Retraction Watch Database[4] with those receiving comments mentioning problematic citations on PubPeer; (3) 7 (2.8%) pairs are from corrections, errata, and corrigenda available in the PubMed database.

Three additional inclusion criteria were applied to the dataset to ensure data quality and experiment reproducibility. First, both the citing and the reference article should have digital versions that are findable through search engines. Second, the reference should be a journal article such that text extraction from the PDF and embedding-based retrieval of text chunks are feasible. Third, the statement to be verified can be uniquely identified in the citing article. The last criterion is necessary because a reference can be cited multiple times for different statements in a paper.

C Prompt Template

User:
You are an experienced scientific writer and editor. You will be given a statement from an article that cites a reference article and information from the reference article. You will determine and explain if the reference article supports the statement.

Specifically, choose a label from "Fully substantiated", "Partially substantiated", and "Unsubstantiated". Further explanations of the labels are as follows:
"Fully substantiated": The reference article fully substantiates the relevant part of the statement from the present article.
"Partially substantiated": According to the reference article, there is a minor error in the statement but the error does not invalidate the purpose of the statement.
"Unsubstantiate": The reference part does not substantiate any part of the statement. This could be because the statement is contradictory to, unrelated to, or simply missing from the reference article.

Format your answer in JSON with two elements– "label" and "explanation". Your explanation should be short and concise.
The citing article
Title: {title}
Statement: {statement}

The reference article
Title: {reference_title}
Abstract: {reference_abstract}
Excerpts: {reference_excerpts}

[3] https://pubpeer.com/
[4] http://retractiondatabase.org/

D Examples of Incorrect Predictions

Abstract and excerpts of the reference were provided in the experiment but omitted here.

Example 1 from GPT-4 Turbo
Title of the citing article:
Snowball Earth climate dynamics and Cryogenian geology-geobiology
Statement:
In East Greenland, East Svalbard, and South China [citation mark], 0.2 to 0.3 km (uncompacted) of fine-grained terrigenous strata appear.
Title of the reference:
Reconstruction of paleo-redox conditions and early sulfur cycling during deposition of the Cryogenian Datangpo Formation in South China
Gold label: Fully substantiated
LLM label: Partially substantiated
LLM explanation:
The reference article focuses on the paleo-redox conditions, sulfur cycling, and the deposition environment of the Datangpo Formation in South China during the Cryogenian period, without mentioning the specific thickness of uncompacted fine-grained terrigenous strata in East Greenland, East Svalbard, and South China. Therefore, it does not substantiate the statement regarding the thickness of terrigenous strata in these locations.

Example 2 from GPT-4 Turbo
Title of the citing article:
Transition metal-catalyzed alkyl-alkyl bond formation: Another dimension in cross-coupling chemistry
Statement:
In early studies, primary alkyl electrophiles were coupled with alkylmagnesium reagents (Grignard reagents) in the presence of transition metal catalysts such as copper [citation mark].
Title of the reference:
New developments in palladium catalysed cross coupling: The coupling of alkyl iodides with alkyl Grignard reagents
Gold label: Fully substantiated
LLM label: Partially substantiated
LLM explanation:
The reference article supports the use of alkyl-magnesium reagents (Grignard reagents) in cross-coupling reactions with alkyl electrophiles, specifically alkyl iodides, in the presence of a transition metal catalyst (palladium). However, it does not mention copper as a catalyst, focusing instead on a palladium-based system. Thus, while the reference article substantiates the general process of alkyl-alkyl bond formation using transition metal catalysts and Grignard reagents, it does not support the specific mention of copper as a catalyst.

Example 3 from GPT-4 Turbo
Title of the citing article:
Solar radiation impact on ferrofluid convection with applying electric field
Statement:
The nanoliquids are a mixture of typical fluids with nano-powders illustrating great range of heat transfer [citation mark].
Title of the reference:
Comprehensive evaluation of flat plate and parabolic dish solar collectors' performance using different operating fluids and MWCNT nanofluid in different climatic conditions
Gold label: Unsubstantiated
LLM label: Fully substantiated
LLM explanation:
The reference article directly supports the statement by demonstrating that the addition of nanoparticles (e.g., MWCNT) to base fluids increases the thermal conductivity of the fluid, thereby illustrating a great range of heat transfer enhancement. This is evidenced by the specific mention of increased thermal conductivity and better performance of solar systems using nanofluids.