# *SciCiteVal*: A Multi-Domain Dataset for Scientific Citation Verification

**Qinyue Liu, Yongxin Zhou, Cyril Labbé**
Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG
38000, Grenoble, France
{qinyue.liu,yongxin.zhou,cyril.labbe}@univ-grenoble-alpes.fr

**Abstract**

Citations are an integral and important part of scientific papers. However, there exist erroneous citations ranging from careless mistakes to deliberate misconduct, and there are currently few studies or benchmark datasets dedicated to automated citation verification. To bridge this gap, we introduce *SciCiteVal*, a novel, manually annotated dataset for citation verification. Each instance in *SciCiteVal* pairs a citation context from a citing paper with the corresponding evidence passage extracted from the full text of the cited source. The dataset features a comprehensive taxonomy, where each citation is annotated as "Correct", "Incorrect", or "Unrelated", with the "Incorrect" category further divided into five fine-grained sub-categories. The completed dataset comprises over 1,000 annotated citations, distributed as 302 "Correct", 302 "Incorrect", and 430 "Unrelated" instances. We establish a benchmark by evaluating different Large Language Models (LLMs), providing baseline performance and a detailed analysis. We release *SciCiteVal* as a resource to support the development of citation verification systems and to facilitate research on evidence-based tasks.

**Keywords:** Citation Verification Dataset, Large Language Models (LLMs), Benchmark

## 1. Introduction

Citations are a fundamental component of scientific communication. They can be used to track the way knowledge spreads and to provide a foundation for new hypotheses (Horbach et al., 2021). However, the integrity of the scientific literature is often compromised when researchers make careless mistakes or, in more severe cases, engage in deliberate citation misconduct. For instance, Glenton and Carlsen (2019) documented citations that misrepresented their original research, while da Silva et al. (2023) identified numerous unrelated citations erroneously linked to their work due to bugs in the Digital Object Identifiers (DOIs) resolution process. A further extreme is the citations generated by paper mills — organizations that sell citaions and authorship in legitimate journals (Abalkina and Bishop, 2023). Such inaccurate citations can lead to the misinterpretation of research findings, distort the original authors' intent, and can even have severe consequences for the scientific ecosystem.

Recently, the use of Large Language Models (LLMs) for information retrieval and text generation (Zhou et al., 2025) has further amplified the importance of this issue, especially since LLMs are known to hallucinate and fabricate seemingly plausible but actually non-existent scientific citations (Agrawal et al., 2024). This trend makes the development of automated systems for verifying scientific claims, particularly citation verification, important. Such systems are essential not only for helping researchers maintain citation integrity but also for verifying the reliability of content generated by LLM within scientific contexts.

While recent studies have advanced our understanding of various citation characteristics, automatic verification of citations remains a challenging task. Existing research has produced several datasets for scientific claim verification (Wadden et al., 2020), which involves retrieving evidence for or against a given scientific claim by searching a large corpus of document, framing it primarily as an information retrieval problem. In contrast, citation verification is an auditing issue that answers the question: "Does a particular citation accurately represent the content of the source it cites?"

As far as we know, few studies have developed dedicated datasets for citation verification or evaluated LLMs on this task. To address this, we introduce *SciCiteVal*, a novel dataset for citation verification. Our dataset contains manual annotations labeling citations as "Correct", "Incorrect" or "Unrelated". For citations labeled "Incorrect", we further define five detailed sub-categories to characterize the nature of the inaccuracy. Each data sample is structured as a pair, consisting of the **citation context** from the citing paper and the corresponding **evidence passage** from the cited paper. In this work, our main contributions are as follows:

- We define a comprehensive taxonomy for citation verification, comprising three major categories ("Correct", "Incorrect", "Unrelated") and a fine-grained set of five sub-categories classifying common types of citation inaccuracies.
- We manually annotated a dataset for the citation verification task, addressing a resource



*Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 1603–1611
11–16 May 2026. 

gap in this domain[1].

- We establish a benchmark by conducting a thorough evaluation of several state-of-the-art Large Language Models (LLMs) on this task, providing baseline performance and analysis.
- For all "Incorrect" citations, we provide simple explanations of *how* the cited source was misrepresented. These rationales can support future work on fine-grained error classification and explainable verification.

## 2. Related Work

### 2.1. Citation Categorization

Some research and tools focus on classifying citations and identifying their function or intent. For example, Liu (2017) uses the labels *"negative", "positive", and "objective"* for citation sentiment analysis; Te et al. (2022) aims to detect citation polarity, using *"critical" and "non-critical"* as citation labels (Bordignon, 2022). Liu et al. (2024b) classifies citations into *"Reliable in domain" and "Erroneous out of domain"*. McIntyre and Haussmann (2021) proposes a more fine-grained classification approach: a four-label classification framework with categories such as *"clearly support", "unsupport", "ambiguous", or "empty"*. These progression lead to the development of tools, such as that by Nicholson et al. (2021), which automate the identification of citation contexts and their functions — such as supporting, mentioning, or contrasting — to analyze how scholarly works are referenced throughout the literature. Our work focuses on deeper semantic categorization, which requires retrieving and reasoning about additional information from the cited papers themselves, going beyond the surface features of the citing sentences.

### 2.2. Datasets about Scientific Articles

In order to construct our dataset, we investigate several datasets about scientific articles. We considered both summarization-focused datasets such as *Scisummnet* (Yasunaga et al., 2019) and *PLOS* (Goldsack et al., 2022), as well as Question Answering (QA)-oriented collections including *QASPER* (Dasigi et al., 2021) and *QASA* (Lee et al., 2023). Among these, the *QASA* dataset stood out for its rigorously constructed and validated question-answer pairs. Furthermore, the structure of *QASA*'s answers can be easily transformed into citation contexts, making it the most appropriate foundation for our dataset.

[1]Our dataset is available at: https://huggingface.co/datasets/birdie0111/SciCiteVal.

### 2.3. Scientific Claim Verification

Scientific claim verification is a similar task to citation verification, which involves retrieving evidence for or against a given scientific claim by searching a large corpus of document. Wadden et al. (2020) introduced *SciFact*, a dataset for scientific claim verification comprising 1.4K expert-written claims paired with evidence from paper abstracts. Each instance is annotated with a label (SUPPORTS or REFUTES) and a rationale. Subsequent work has employed similar tasks with varied label sets; for instance, Alvarez et al. (2024) introduced *SCitance*, a dataset derived from *SciFact*, which classifies citations as supporting, contradicting, or unrelated to a claim. A key limitation of these prior works is their reliance on a narrow evidence base, typically restricted to paper abstracts or tables (Ho et al., 2025). In contrast, our dataset provides a more comprehensive set of evidence, including not only abstract information but also the main text of the papers, enabling a more robust and realistic citation verification process. Furthermore, our dataset covers multiple domains, providing broader insights than previous single-domain benchmarks.

### 2.4. Citation Verification Tasks

Similar to our citation verification task, there are also recent works that explore citation integrity. For example, Liu et al. (2024a) worked on detecting "anomalous citations" which they define as suspicious citations that are used to artificially enhance the impact of authors & publications and hack the impact factor of journals, rather than pointing to the knowledge that contributes to the citing work. Their approach mainly makes use of authorship networks and abstracts of the papers, different from ours which depends directly on the semantic correctness between citation context and its justifying cited content. The work of Sarol et al. (2024) is the closest to our task, they annotated 3,063 citations from 100 highly cited biomedical papers, and proposed NLP methods to identify erroneous citations. They also proposed a taxonomy for their *NOT_ACCURATE* citations.

Our work differs from theirs in that we define categories and subcategories slightly differently. Our dataset primarily consists of citations from social science, biology and machine learning fields, while theirs is based in the biomedical field. Our primary goal is to establish a benchmark specifically designed to evaluate models on the challenging task of citation verification. However, we see the potential to expand our dataset based on theirs in the future.

## 3. Dataset Creation

Our dataset contains 1,034 citations distributed across three categories: 302 Correct citations, 302 Incorrect citations, and 430 Unrelated citations. The citations are from scientific papers in social science, biology and machine learning domains. Both Correct and Incorrect citations are adapted from the QASA dataset (Lee et al., 2023), whereas unrelated citations are extracted from real citations that cite unrelated articles (Liu et al., 2025).

The category definitions are listed below, and Table 1 provides representative examples for each.

- **Correct:** Citation contexts that accurately reference the justifying content within the cited article.
- **Incorrect:** Citation contexts that misrepresent the cited article while still being related to the cited article.
- **Unrelated:** Citation contexts that incorrectly cite articles and are unrelated to the cited article.

### 3.1. Base Datasets: QASA

The construction of our citation verification dataset builds partially upon the QASA dataset, a choice justified in Section 2.2.

The QASA dataset contains 1,798 QA pairs from AI/ML papers, with questions posed by readers and answers provided by experts (denoted as "composition"). Annotators also identified supporting evidence paragraphs within the papers (denoted as "evidence"). To validate quality, domain experts manually assessed 100 randomly sampled questions, confirming that 90% of answers were correct and 87% were well-grounded. Consequently, we repurpose and modify the "composition" as citation contexts and the "evidence" as cited content. We exclusively utilize "shallow" and "testing" questions from their "answerable questions" subset, as their "compositions" are more readily justifiable with cited content, ensuring greater accuracy in our subsequent modifications.

### 3.2. Our Dataset: SciCiteVal

Our dataset mainly consists of two files, one refined for making experiments (experiment file), and the other containing more annotation details (annotation file). The experiment file is mainly composed of four columns: "Citation_context", "Cited_content", "Label" and "Twist_category". To ensure traceability, we keep the original QASA dataset IDs for all derived citations, and provide bref explanations on how the citation is distorted in our annotation file. The "Twist_category" column specifically annotates the sub-category of each "Incorrect" citation.

To collect the "Correct" citations, we manually verify that each "composition" from the QASA dataset can function as a citation context written to answer the corresponding question. We exclude composition types containing only "Yes" or "No" answers, and ensure that the justifying content is present in the associated "evidence" section from Qasa dataset, which will be seen as "Cited_content" in our dataset. The validated "composition" are noted as the citation context under the "Citation_context" column.

For "Incorrect" citations, two annotators with NLP backgrounds read the original "composition" and systematically distort it to form a citation context that inaccurately references the "evidence" section according to our given taxonomy. The annotators also make minimal necessary modifications to enhance the fluency of the compositions as citation contexts. The detailed taxonomy of these alterations is provided in Section 3.3. We also require annotators to write a brief explanation of their distortions in the "Twist_reason" section in our annotation file.

The "Unrelated" citations are extracted from our previous work (Liu et al., 2025), covering scientific papers across diverse fields, such as social science, biology and machine learning. All collected citation contexts are complete sentences from the main body of texts, excluding any from tables or figures, and the abstracts of the cited articles are seen as the "Cited_content" in this case. The dataset is annotated by multiple annotators. The majority of this category are sourced from various papers within Vickers's case (da Silva et al., 2023).

### 3.3. Detailed Taxonomy for Incorrect Citations

To the best of our knowledge, research on the taxonomy of "Incorrect" citations remains limited, and collecting real samples of such citations is resource-intensive and time-consuming. Consequently, we synthesize such citations by strategically distorting correct ones. Our taxonomy is developed through a preliminary study: two annotators freely distorted 20 random compositions, documenting their distortion strategies. We then integrate these observations to synthesize a set of distinct sub-categories, defined by the common distortion patterns applied. This process yielded a final taxonomy of five sub-categories for "Incorrect" citations.

The definitions and examples for each subcategory are provided in Table 2. Notably, the "Tortured Phrases" category is inspired by Cabanac et al. (2021). We draw from their established list

| Category | Citation Context | Cited Content |
|---|---|---|
| Correct | Using lexical matching makes it difficult to identify synonyms or to distinguish between ambiguous words. | Information Retrieval (IR) is a central component of many natural language applications. Traditionally, lexical methods (Robertson et al., 1994) have been used to search through text content. However, these methods suffer from the lexical gap (Berger et al., 2000) and are not able to recognize synonyms and distinguish between ambiguous words. |
| Incorrect | During indexing, they use another server with the same CPU and system memory specifications but which *has two Titan V GPUs attached, each with 8 GiBs of memory.* Across all experiments, only one GPU is dedicated per query for retrieval (i. , for methods with neural computations) but we use up to all four GPUs during indexing. | To evaluate the latency of neural re-ranking models in §4.2, we use a single Tesla V100 GPU that has 32 GiBs of memory on a server with two Intel Xeon Gold 6132 CPUs, each with 14 physical cores (24 hyperthreads), and 469 GiBs of RAM. For the mostly CPU-based retrieval experiments in §4.3 and the indexing experiments in §4.5, we use another server with the same CPU and system memory specifications but which *has four Titan V GPUs attached, each with 12 GiBs of memory.* Across all experiments, only one GPU is dedicated per query for retrieval (i.e., for methods with neural computations) but we use up to all four GPUs during indexing. |
| Unrelated | HSG involves the actions and means adopted by society to organize to promote and protect the health of its population. | Male moths compete to arrive first at a female releasing pheromone. A new study reveals that additional pheromone cues released only by younger females may prompt males to avoid them in favor of older but more fecund females. |

Table 1: Samples of each category in our dataset.

of tortured phrases and systematically incorporate them into compositions to create this specific type of distortion.

# 4. Experiments

## 4.1. Task Overview

We evaluate several Large Language Models (LLMs) on the citation verification task. For each sample, we provide the model with a prompt containing the citation context and the corresponding cited content (for unrelated citations, we use the article's abstract). The prompt instructs the model to classify the citation into one of three categories: "Correct", "Incorrect", or "Unrelated". We do not require the models to predict the fine-grained subcategories of "Incorrect" citations.

## 4.2. Models

We select several state-of-the-art LLMs of varying sizes known for their strong semantic understanding capabilities. The chosen models include Qwen3-14B[2] (QwenTeam, 2025), Qwen3-8B[3] (QwenTeam, 2025), Mistral-7B-Instruct[4] (Joren et al., 2025), Mistral-small-24B-Instruct[5] (MistralAI, 2025), Llama3.1-8B-Instruct[6] (AI@Meta, 2024), and Deepseek-chat[7] (DeepSeek-AI, 2025).

## 4.3. Experimental Setup

### 4.3.1. Zero-shot Prompting

For all models, we use an identical prompt structure. It comprises the definitions of the three labels, the citation context with its cited content, and a clear instruction to analyze the relationship and output the correct label. The prompt is formulated as follows in Prompt 4.3.1.

### 4.3.2. Few-shot prompting

In the few-shot prompting setting, we enhance the prompt by providing one example from each citation category (three examples in total) before the label definitions. We also append an explicit instruction that constrains the model's output to only the three predefined categories. The rest of the prompt is identical to the zero-shot version. The final prompt is structured as follows in Prompt 4.3.2.

[2] https://huggingface.co/Qwen/Qwen3-14B

[3] https://huggingface.co/Qwen/Qwen3-8B

[4] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3

[5] https://huggingface.co/mistralai/Mistral-Small-24B-Instruct-2501

[6] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[7] https://huggingface.co/deepseek-ai/DeepSeek-V3.2-Exp

| Sub-category | Definition | Example |
|---|---|---|
| Opposite | Reverse the semantic polarity of the original phrase. | "Can" →"Cannot", "Has an impact" →"Does not have an impact" |
| Change Number | Modify key numerical values mentioned in the text. | "The accuracy is 80%" →"The accuracy is 95%", "We investigated 300 samples" →"We investigated 500 samples |
| Add Unmentioned Information | Add new information or claims that are absent from the original content. | "We applied two approaches: 1..., 2..." →"We applied three approaches: 1..., 2..., 3..." |
| Tortured Phrases | Replace a standard scientific term with an unconventional, often nonsensical phrase. | "artificial intelligence" →"counterfeit consciousness" |
| Change Concept | Change a core concept, potentially leading to a significant semantic shift. | "A certain part" →"The entire part", "reinforcement learning" →"supervised learning", or swapping definitions |

Table 2: Definitions of sub-categories for "Incorrect" citations.

**Zero-shot Prompt**

**System:**
Please analyze the given "citation context" and the "cited content" carefully and then verify if the "citation context" correctly cites the "cited content".

Only output:
- **Correct** — if the citation context correctly cites the cited content;
- **Incorrect** — if the citation context is incorrect but still related to the cited content (even a subtle fault counts);
- **Unrelated** — if the citation context is incorrect and also unrelated to the cited content.

---

**User:**
The citation context is: `(The given citation context)`
The cited content is: `(The given cited content)`

**Few-shot Prompt**

**System:**

Please analyze the given "citation context" and the "cited content" carefully and then verify if the "citation context" correctly cites the "cited content".

An example for "Correct" citation is like:
The citation context: (Citation context of the example)
The cited content: (Cited content of the example).

An example for "Incorrect" citation is like:
The citation context: (Citation context of the example)
The cited content: (Cited content of the example).

An example for "Unrelated" citation is like:
The citation context: (Citation context of the example)
The cited content: (Cited content of the example).

Only output "Correct" if the citation context correctly cites the cited content;
"Incorrect" if the citation context is incorrect but still related to the cited content (Even a subtle fault counts);
"Unrelated" if the citation context is incorrect and also unrelated to the cited content.
Do not output anything else.

---

**User:**

The citation context is: `(The given citation context)`,
The cited content is: `(The given cited content)`

### 4.4. Evaluation Metrics

We first report and compare the overall accuracy and F1 scores, as well as the results between the zero-shot and few-shot prompting setups. Given the varying difficulty of verifying each citation category, we also evaluate model performance using a confusion matrix.

## 5. Results and Analysis

### 5.1. General Analysis

Table 3 presents the overall evaluation results. On our dataset, `Deepseek-chat` achieves the highest accuracy under both zero-shot and few-shot settings. For F1 score, `Qwen3-14B` ranks first in the zero-shot condition, while `Deepseek-chat` obtains the best F1 score with few-shot prompting.

A key observation is that most models struggle to distinguish between "Correct" and "Incorrect" citations, exhibiting a pronounced bias towards one of these two classes. Although this bias is moder-

| Model | Zero-shot Result | | | | | Few-shot Result | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | C | I | U | A | F1 | C | I | U | A | F1 |
| Deepseek-chat | 0.74 | 0.63 | 0.94 | **0.79** | 0.67 | 0.78 | 0.60 | 0.99 | **0.81** | **0.71** |
| Mistral-7B | **0.94** | 0.08 | 0.99 | 0.71 | 0.66 | **0.97** | 0.05 | **1.00** | 0.71 | 0.66 |
| Mistral-small-24B | 0.01 | **1.00** | 0.77 | 0.62 | 0.02 | 0.35 | 0.78 | **1.00** | 0.75 | 0.44 |
| Qwen3-8B | 0.73 | 0.48 | 0.87 | 0.72 | 0.60 | 0.63 | 0.51 | **1.00** | 0.75 | 0.59 |
| Qwen3-14B | 0.85 | 0.34 | **1.00** | 0.76 | **0.68** | 0.87 | 0.32 | **1.00** | 0.76 | 0.68 |
| Llama3.1-8B | 0.16 | 0.89 | 0.97 | 0.71 | 0.24 | 0.23 | **0.86** | 0.85 | 0.67 | 0.29 |

Table 3: Performance of models on our dataset. Columns **C**, **I**, and **U** represent categorical accuracy for "Correct", "Incorrect", and "Unrelated" citations respectively, while **A** (Overall Accuracy) and **F1** (F1 Score) columns with gray background indicate overall performance metrics. The best results are in bold.

ately reduced with few-shot prompting, detecting "Incorrect" citations remains a challenging task for all models.

Furthermore, we observe divergent behaviors among models from the same family. `Mistral-7B` tends to misclassify "Incorrect" citations as "Correct", whereas `Mistral-small-24B` exhibits the opposite tendency. After applying few-shot prompting, `Mistral-small-24B` demonstrates substantial improvement in classifying both "Correct" and "Unrelated" citations, leading to a notable increase in its overall accuracy and F1 score.

For most models, few-shot prompting enhances the classification performance for "Unrelated" and "Correct" citations, but leads to a drop in the ability to identify "Incorrect" ones. This trade-off is evident in the confusion matrices. For instance, `Mistral-small-24B` becomes less strict in judging "Incorrect" citations after few-shot learning (comparing Figure 1a and Figure 1c), as it reclassifies more number of them as "Correct". Conversely, `Qwen3-8B` exhibits the opposite tendency (Figure 1b vs. Figure 1d), showing a heightened strictness after few-shot learning that causes more instances from "Correct" category to be misclassified as "Incorrect".

## 5.2. Unwanted Output Analysis

During experimentation, we observe that several models generate reasoning or explanatory text alongside the category label, despite explicit instructions in the prompt to output only the three specified categories. For example, `Llama3.1-8B` explains every single output with its analysis of contexts and explanation for justifying its output category with zero-shot prompt, while `Mistral-7B` sometimes output its reasoning after analyzing contexts, which is mostly clear, but without indicating which category to classify. For these instances, we manually assign a label based on the model's analysis and reasoning. To address these non-compliant outputs, we introduce an additional phrase in the few-shot prompt explicitly prohibiting any extraneous output. This modification successfully increases the rate of compliant outputs across all models. A detailed statistics of this analysis is provided in Table 4.

| Model | Zero-shot | Few-shot |
|---|---|---|
| Deepseek-chat | 1 | 0 |
| Mistral-7B | 115 | 128 |
| Mistral-small-24B | 654 | 0 |
| Qwen3-8B | 10 | 0 |
| Qwen3-14B | 27 | 0 |
| Llama3.1-8B | 1034 | 74 |

Table 4: Number of unwanted output from different models (1034 in Total).

| Sub-category | Found | Total | Ratio |
|---|---|---|---|
| Opposite | 49 | 58 | 0.84 |
| Change Number | 34 | 42 | 0.81 |
| Add Unmentioned Info | 7 | 20 | 0.35 |
| Tortured Phrases | 5 | 39 | 0.13 |
| Change Concept | 86 | 143 | 0.60 |

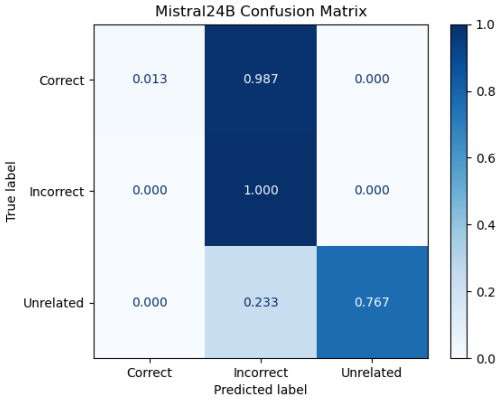
Table 5: Number of *Incorrect* Citations found by Deepseek-chat (grouped by sub-category).

## 5.3. *Incorrect* Citations Output Analysis

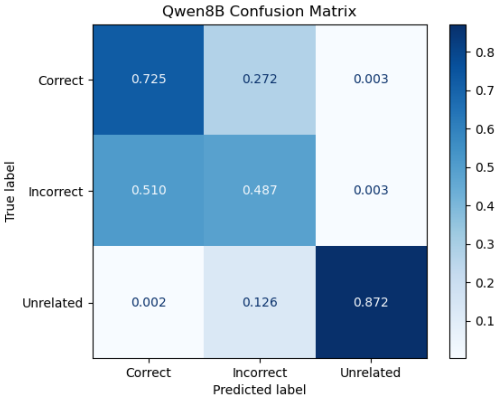
We perform a fine-grained analysis of the "Incorrect" citation category to determine which sub-categories are more readily identified as "Incorrect" by the models. We examine the predictions from the top-performing `Deepseek-chat` model with few-shot prompting. The results, detailed in Table 5, indicate that the model struggles to identify "Incorrect" citations from the "Add Unmentioned Information" (35% identified) and "Tortured Phrases" (12.8% identified) sub-categories. In contrast, it demonstrates relative proficiency in classifying "Incorrect" citations from the "Opposite" and "Change Number" sub-categories.

# 6. Conclusion

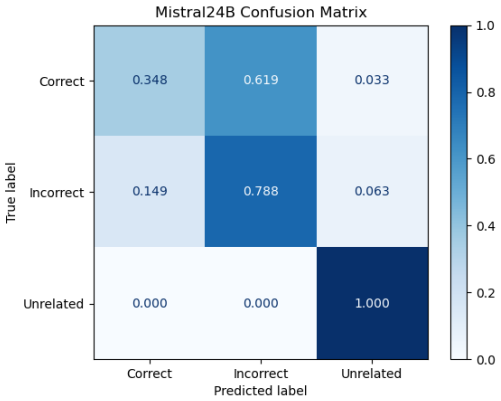
In this paper, we present a manually annotated dataset designed for citation verification, which is



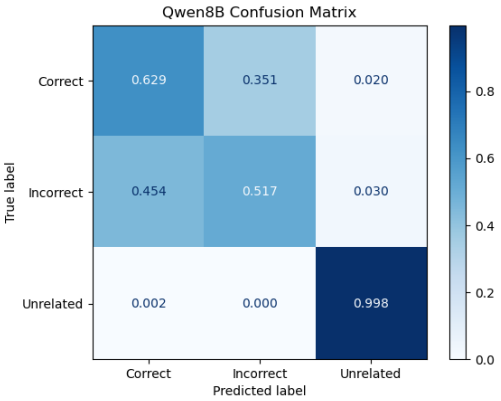

(a) Zero-shot Confusion Matrix for Mistral-Small-24B.




(b) Zero-shot Confusion Matrix for Qwen3-8B.




(c) Few-shot Confusion Matrix for Mistral-Small-24B.




(d) Few-shot Confusion Matrix for Qwen3-8B.

Figure 1: Confusion matrices comparing zero-shot and few-shot performance of `Mistral-Small-24B` and `Qwen3-8B` models: (a) Zero-shot `Mistral-Small-24B`, (b) Zero-shot `Qwen3-8B`, (c) Few-shot `Mistral-Small-24B`, (d) Few-shot `Qwen3-8B`.

an important task in maintaining scientific integrity. Our dataset provides structured pairs of citation context and full-text evidence, annotated using a taxonomy that categorizes citations as "Correct", "Incorrect", or "Unrelated". The "Incorrect" category is further divided into five fine-grained subcategories to enable detailed analysis.

Our benchmarking, which evaluates six models from four families, reveals a clear performance hierarchy. While `Deepseek-chat` achieves state-of-the-art performance (0.79 zero-shot, 0.81 few-shot), our dataset presents a significant challenge for other models. For instance, `Mistral-small-24B` and `Llama3.1-8B` achieve accuracies of only 0.62 (zero-shot) and 0.67 (few-shot), respectively. Furthermore, even the top-performing model, `DeepSeek-Chat`, exhibits difficulty in identifying "Incorrect" citations, particularly for categories like "Tortured phrases" and "Add Unmentioned Information" — highlighting a lack of the precise reasoning required for robust citation verification. Smaller models also demonstrate significant prediction bias on our dataset.

For future work, we plan to create a larger and more diverse dataset, which is essential for training robust models. We will also deepen the semantic analysis of citation errors to refine the taxonomy, directly addressing the model weaknesses identified in this study. In the longer term, we will focus on three key goals: (1) refine the taxonomy to be more detailed, and expand our dataset with more human-annotated data from more diverse sources. (2) benchmark a wider range of models on the citation classification task under the future fine-grained taxonomy, (3) incorporate more constructed human-annotated rationales into the expanded dataset to support additional NLP tasks that require explanatory reasoning, (4) explore optimization strategies to enhance model performance on citation verification task, and (5) contributing our dataset and findings to the broader research community to assist in maintaining citation integrity.

## Limitations

Our work presents several limitations that offer pathways for future research. Firstly, the scale of our dataset is currently limited, as 58% of it is drawn from and extends the single QASA dataset. Its breadth and generalizability could be strengthened by incorporating and adapting instances from other scientific corpora. Secondly, while our taxonomy for "Incorrect" citations provides a useful framework, its robustness could be further enhanced through a more comprehensive analysis of real-world citation error cases. Finally, our benchmarking, while informative, was conducted on a select set of models. It would be highly insightful for future work to evaluate the performance of more advanced proprietary models, such as `GPT-5` and `Claude-Sonnet`, on our dataset to gain a more comprehensive understanding of state-of-the-art capabilities.

## Acknowledgments

We acknowledge the NanoBubbles project that has received Synergy grant funding from the European Research Council (ERC), within the European Union's Horizon 2020 program, grant agreement no. 951393. This work was also partially supported by MIAI@Grenoble-Alpes (ANR-23-IACL-0006), and partially supported by the "Intelligent Systems for Bridging Data, Knowledge and Humans" axis of the Grenoble Computer Science Laboratory (LIG).

## Bibliographical References

Anna Abalkina and Dorothy Bishop. 2023. Paper mills: a novel form of publishing malpractice affecting psychology. *Meta-Psychology*, 7.

Ayush Agrawal, Mirac Suzgun, Lester Mackey, and Adam Kalai. 2024. Do language models know when they're hallucinating references? In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 912–928, St. Julian's, Malta. Association for Computational Linguistics.

AI@Meta. 2024. The llama 3 herd of models.

Frederique Bordignon. 2022. Critical citations in knowledge construction and citation analysis: from paradox to definition. *Scientometrics*, 127(2):959–972.

Guillaume Cabanac, Cyril Labbé, and Alexander Magazinov. 2021. Tortured phrases: A dubious writing style emerging in science. evidence of critical issues affecting established journals.

Jaime A. Teixeira da Silva, Neil J. Vickers, and Serhii Nazarovets. 2023. From citation metrics to citation ethics: Critical examination of a highly-cited 2017 moth pheromone paper. *Scientometrics*, 129(1):693–703.

DeepSeek-AI. 2025. Deepseek-v3.2-exp: Boosting long-context efficiency with deepseek sparse attention.

Claire Glenton and Benedicte Carlsen. 2019. When "normal" becomes normative: A case study of researchers' quotation errors when referring to a focus group sample size study. *International Journal of Qualitative Methods*, 18:1609406919841251.

Serge Horbach, Kaare Aagaard, and Jesper W. Schneider. 2021. Meta-Research: How problematic citing practices distort science. MetaArXiv aqyhg, Center for Open Science.

Hailey Joren, Jianyi Zhang, Chun-Sung Ferng, Da-Cheng Juan, Ankur Taly, and Cyrus Rashtchian. 2025. Sufficient context: A new lens on retrieval augmented generation systems.

Haixia Liu. 2017. Sentiment analysis of citations using word2vec. *CoRR*, abs/1704.00177.

Jiaying Liu, Xiaomei Bai, Mengying Wang, Suppawong Tuarob, and Feng Xia. 2024a. Anomalous citations detection in academic networks. *Artificial Intelligence Review*, 57(4).

Qinyue Liu, Amira Barhoumi, and Cyril Labbé. 2024b. Miscitations in scientific papers: Dataset and detection. pages 53–65.

Trevor McIntyre and N.S. Haussmann. 2021. Declining citation accuracy in polar research. *Polar Record*, 57:e43.

Mistral-AI. 2025. Magistral.

Josh M. Nicholson, Milo Mordaunt, Patrice Lopez, Ashish Uppala, Domenic Rosati, Neves P. Rodrigues, Peter Grabitz, and Sean C. Rife. 2021. scite: A smart citation index that displays the context of citations and classifies their intent using deep learning. *Quantitative Science Studies*, 2(3):882–898.

QwenTeam. 2025. Qwen3 technical report.

Maria Janina Sarol, Shufan Ming, Shruthan Radhakrishna, Jodi Schneider, and Halil Kilicoglu. 2024. Assessing citation integrity in biomedical publications: corpus annotation and nlp models. *Bioinformatics*, 40(7):btae420.

Sonita Te, Amira Barhoumi, Martin Lentschat, Frédérique Bordignon, Cyril Labbé, and François Portet. 2022. *Citation Context Classification: Critical vs Non-critical*. Association for Computational Linguistics, Gyeongju, Republic of Korea.

Yongxin Zhou, Philippe Mulhem, and Didier Schwab. 2025. Tempperturb-eval: On the joint effects of internal temperature and external perturbations in rag robustness.

## 7. Language Resource References

Carlos Alvarez, Maxwell Bennett, and Lucy Wang. 2024. Zero-shot scientific claim verification using LLMs and citation text. In *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*, pages 269–276, Bangkok, Thailand. Association for Computational Linguistics.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers.

Tomas Goldsack, Zhihao Zhang, Chenghua Lin, and Carolina Scarton. 2022. Making science simple: Corpora for the lay summarisation of scientific literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10589–10604, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xanh Ho, Sunisth Kumar, Yun-Ang Wu, Florian Boudin, Atsuhiro Takasu, and Akiko Aizawa. 2025. Table-text alignment: Explaining claim verification against tables in scientific papers.

Yoonjoo Lee, Kyungjae Lee, Sunghyun Park, Dasol Hwang, Jaehyeon Kim, Hong-in Lee, and Moontae Lee. 2023. Qasa: Advanced question answering on scientific articles. In *Proceedings of the 40th International Conference on Machine Learning*.

Qinyue Liu, Yağmur Öztürk, Tiziri Terkmani, François Portet, and Cyril Labbé. 2025. Citescreener: A pipeline for citation verification in digital libraries with datasets. pages 158–166.

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, Online. Association for Computational Linguistics.

Michihiro Yasunaga, Jungo Kasai, Rui Zhang, Alexander R. Fabbri, Irene Li, Dan Friedman, and Dragomir R. Radev. 2019. Scisummnet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks.