CiteGuard: Faithful Citation Attribution for LLMs via Retrieval-Augmented Validation

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have emerged as promising assistants for scientific writing. However, there have been concerns regarding the quality and reliability of the generated text, one of which is the citation accuracy and faithfulness. While most recent work relies on methods such as LLM-as-a-Judge, the reliability of LLM-as-a-Judge alone is also in doubt. In this work, we reframe citation evaluation as a problem of citation attribution alignment, which assesses whether LLM-generated citations match those a human author would include for the same text. We propose *CiteGuard*, a retrieval-aware agent framework designed to provide more faithful grounding for citation validation. *CiteGuard* improves the prior baseline by 17%, and achieves up to 68.1% accuracy on the CiteME benchmark, approaching human-level performance (69.7%). It also enables the identification of alternative but valid citations and demonstrates generalization ability for cross-domain citation attribution.[1]

1 Introduction

"*If I have seen further than others, it is by standing upon the shoulders of giants*" — Isaac Newton.

Scientific research often progresses by building on the foundation of prior knowledge. Therefore, a thorough and faithful literature review and citation attribution of claims are essential to understand the history and scope of a subject area, and ensure that new findings are properly contextualized (Salton and Bergmark, 1979; Snyder, 2019; Chigbu et al., 2023). However, conducting such practices has been increasingly difficult due to the rapid growth in the number of scientific publications (Larsen and Von Ins, 2010; Bornmann and Mutz, 2015). Recently, over 50 citation hallucinations were found in 300 ICLR 2026 submissions (Shmatko et al., 2025).

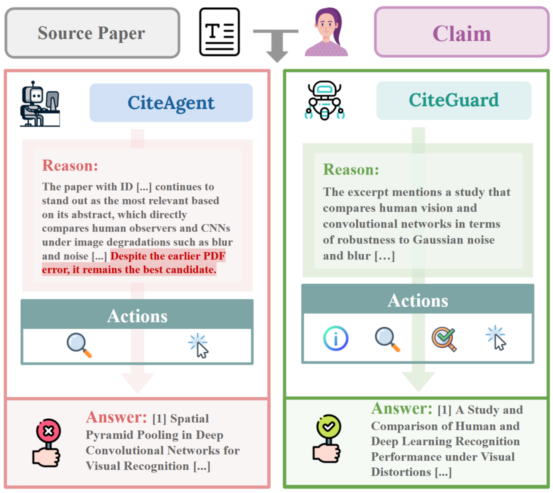

Figure 1: *CiteGuard* succeeds through expanded retrieval actions, whereas CiteAgent (Press et al., 2024) fails due to OpenPDF access error.

Method	Precision	Recall	F1
Zero-shot abstract	1.0	0.17	0.29
Few-shot abstract	1.0	0.16	0.27
Zero-shot full text	1.0	0.36	0.53
Few-shot full text	1.0	0.38	0.55

Table 1: ChatGPT-4o accuracy on citation attribution in the CiteME benchmark.

Large Language Models (LLMs) and LLM agents have emerged as potentially useful tools to alleviate the burden of researchers and support scientific writing (Lu et al., 2024; Yamada et al., 2025; Asai et al., 2024a; Wang et al., 2025). One of the main concerns is hallucinations in LLM (Ji et al., 2023; Huang et al., 2025). For instance, LLMs can generate up to 78-90% fabricated citations (Asai et al., 2024a) and misattribute findings to incorrect sources (Walters and Wilder, 2023).

Retrieval-augmented generation (Lewis et al., 2020; Fan et al., 2025) has been proposed to miti-

[1] Our code is available at https://anonymous.4open.science/r/CiteGuard-FCDC.

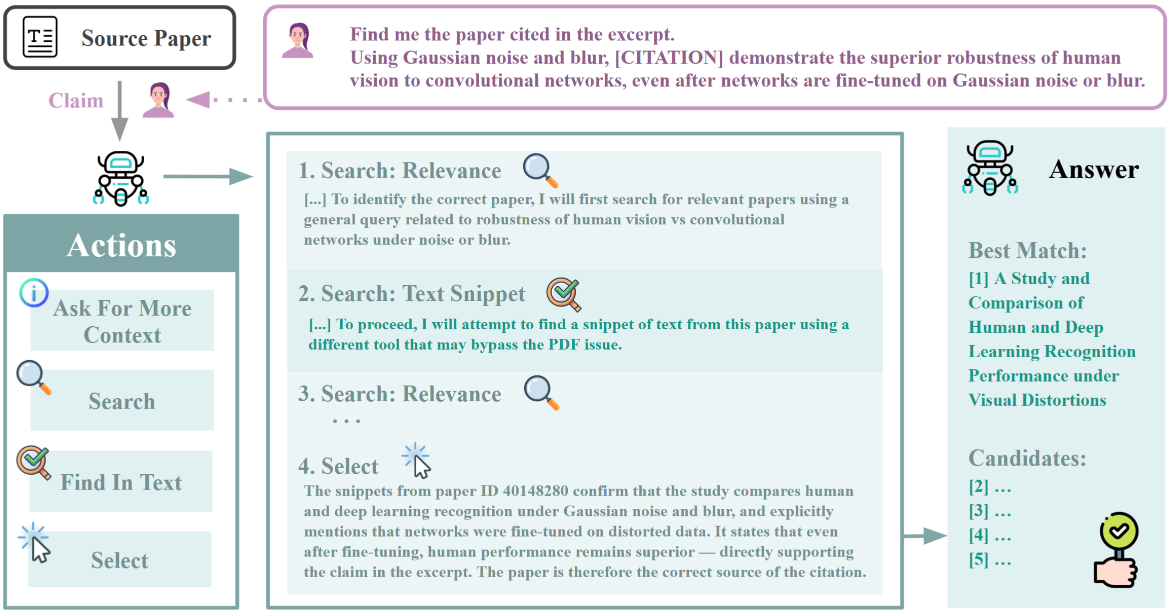

Figure 2: *CiteGuard* takes an excerpt and performs searches to find a paper that best matches the missing citation.

gate hallucinations in LLM by retrieving external knowledge to validate the generated text during training data preparation or at inference time (Wang et al., 2024b; Asai et al., 2024a; Wang et al., 2024a, 2025). LLM-as-a-Judge is often used to prepare training data (Asai et al., 2024a,b) or to evaluate generated text (Asai et al., 2024a; Wang et al., 2024b; He et al., 2025) as it is more scalable in practice, despite the risk of bias and overdependence on LLMs' capabilities (Ye et al., 2024; Thakur et al., 2024). LLM-as-a-Judge often assumes that the retrieved knowledge used for the generation is available, limiting the use case to evaluating retrieval-augmented output. Furthermore, it does not account for situations where the evaluation requires grounding (Krumdick et al., 2025), such as broader textual understanding, cross-referencing multiple sources, or interpreting ambiguous claims.

We conduct an evaluation of the reliability of LLM-as-a-Judge for citation attribution of human-written scientific claims and their references. Although LLMs can recognize apparently incorrect citations, they often reject correct citations due to limited domain-specific knowledge, resulting in a recall as low as 16-17% (Table 1). For instance, LLMs are are sensitive to minor variations in terminology (example in App. E). This could potentially lead to incorrect evaluation of existing methods and limit the performance of trained LLMs when the training data are filtered using LLM-as-a-Judge.

We propose *CiteGuard*, an agent that provides more faithful and generalizable citation attribution through retrieval-augmented validation. Prior work, CiteAgent (Press et al., 2024) aims to accurately cite scientific claims, although achieving accuracy higher than direct prompting, CiteAgent's accuracy (35.3%), still falls short of human performance. We propose additional tools (i.e, to search for the context of the scientific claim and to perform a more robust search for paper content) and resulting in a +17% accuracy over CiteAgent under the same settings. When paired with Deepseek-R1 (Guo et al., 2025), *CiteGuard* can achieve performance (68.1%) which approaches that of a human (69.7%). Human evaluation indicates that *CiteGuard* can suggest additional citations that were missed by the original benchmark. Exploratory experiments show that *CiteGuard* has the potential to generalize to cross-domain and long paragraph settings. Our contributions are:

- We propose *CiteGuard*, an agent that provides faithful citation attribution by suggesting multiple appropriate references.
- We conduct a detailed analysis and collect human annotations of alternative citations that are not captured by the current CiteME benchmark. We also collect 30 new samples to extend the benchmark to cover cross-domain (biomedical) and long paragraph scenarios, denoted as **CiteMulti**.
- We conduct experiments to show that *CiteGuard* significantly improves accuracy in finding the correct reference, suggesting relevant alternative citations, and that these abilities have the potential to be generalized to cross-domain and long paragraph scenarios.

2 CiteGuard

2.1 Problem Formulation

We formulate the task of finding reference(s) for N excerpts $x_1, x_2, ..., x_N$ given a pool of n possible reference candidates $r_1, r_2, ...r_n$. We have a ground-truth labeling function $y(x_i)$ that can map any excerpt x_i to a ground-truth reference $r^* : y(x_i) = r^*$. We also have another labeling function $\hat{y}(x_i)$ from human annotations that can map any excerpt x_i to a set of k ground truth references $\hat{r}^* = \hat{r_1}^*, ..., \hat{r}_k^*$: $\hat{y}(x_i) = \hat{r}^*$. This is different from the CiteME (Press et al., 2024) setting, where there is only one ground truth reference.

The goal of *CiteGuard* is to find a mapping function f_θ such that $f_\theta(x_i) \approx y(x_i), \forall i = 1, ..., N$.

The accuracy is defined as:

$$Acc(f_\theta) = \frac{1}{N}\sum_{i=1}^{N} 1[f_\theta(x_i) = y(x_i)] \quad (1)$$

The agreement is defined as:

$$Agree(f_\theta) = \frac{1}{N}\sum_{i=1}^{N} 1[f_\theta(x_i) \cap \hat{y}(x_i) \neq \emptyset] \quad (2)$$

2.2 Reference Retrieval

To obtain f_θ, *CiteGuard* introduces new actions in addition to CiteAgent (Press et al., 2024). We provide the set of actions below (examples and prompts used can be found in App. A). These actions are executed in a paper database D (i.e., Semantic Scholar), which we can query using a search query q, and the search result will be appended to R. Each paper $P \in D$ contains a title and abstract content $t \in P$, and a body content, with text snippets denoted as $p_i \in P, \forall i$. The source paper that contains the excerpt is S. We present some of the examples for the actions in Figure 3.

1. **(search_)citation_count/relevance** (adopted): Search for a query in the title and abstract fields, then sort the results by citation count/relevance, defined as

$$\text{Search}_c(q, D) = \text{argsort}_{P \in D}(count(t))$$
$$\text{Search}_r(q, D) = \text{argsort}_{P \in D}(rel(q, t))$$

2. **select** (adopted): Select a paper from the search results, defined as

$$\text{Select}(P \in R)$$

3. **find_in_text**: Search for a query string within the full text of a specified paper, defined as

$$\text{Search}_t(q, P) = \text{argsort}_{p \in P}(rel(q, p))$$

4. **ask_for_more_context**: Retrieve the context for an excerpt from the source paper, defined as

$$\text{Search}_{cont}(q_i, S) = \{q_{i-3}, ..., q_{i+3}\}, q_i \in S$$

5. **search_text_snippet**: Search for a query string in the full text of papers, defined as

$$\text{Search}_{sni}(q, D) = \text{argsort}_{p \in P, P \in D}(rel(q, p))$$

2.3 Iterative Retrieval

Apart from finding only one reference, *CiteGuard* can suggest multiple references when appropriate to provide a better understanding of the current literature and facilitate comparative analysis. Every run of *CiteGuard* suggests one appropriate reference, with subsequent runs searching for a new appropriate reference. A researcher using this agent can manually audit this iterative process and decide when to stop or allow the agent to make the decision.

Let $A_k = \{P^{(1)}, \ldots, P^{(k)}\}$ denote the set of papers selected during k iterations, and we define the exclusion set $E_k := A_k$. At iteration k, the searchable database is $D_k := D \setminus E_{k-1}$, i.e., all retrieval actions that search over papers operate on D_k. Concretely, we use the filtered versions of the search actions:

$$\text{Search}_c^{\neg E}(q, D) = \text{argsort}_{P \in D \setminus E}\big(count(t)\big),$$
$$\text{Search}_r^{\neg E}(q, D) = \text{argsort}_{P \in D \setminus E}\big(rel(q, t)\big),$$
$$\text{Search}_{sni}^{\neg E}(q, D) = \text{argsort}_{p \in P,\ P \in D \setminus E}\big(rel(q, p)\big).$$

Thus, once a paper $P^{(k)}$ is selected, it is excluded from all subsequent search_citation_count, search_relevance, and search_text_snippet actions by updating $E_k \leftarrow E_{k-1} \cup \{P^{(k)}\}$.

3 Experiments

3.1 Benchmarking Citation Attribution for Computer Science Excerpts

We evaluate *CiteGuard* on CiteME (Press et al., 2024), which contains 130 excerpts collected from human-written manuscripts in different Computer

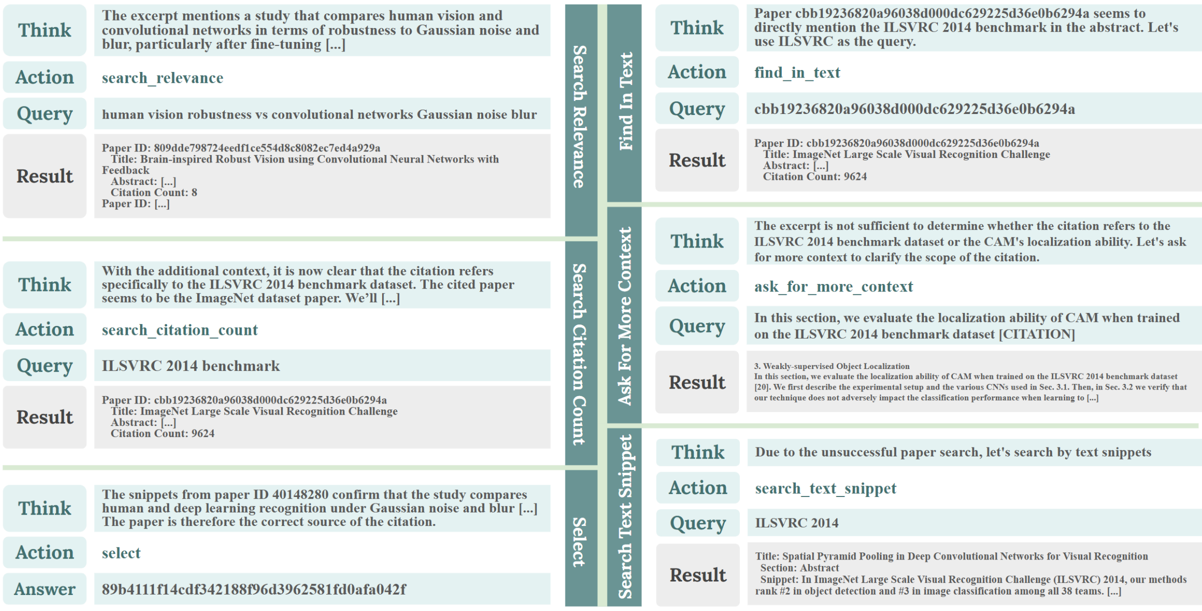

Figure 3: Retrieving Actions. We define six retrieval actions to ensure the efficiency and accuracy of CiteGuard.

	Easy(%)	Medium(%)	Med-Hard(%)	Hard(%)	All(%) ± σ	Agree(%)
CiteAgent+GPT-4o	-	-	-	-	35.3*	-
CiteGuard+GPT-4o	86.4	65.2	20.5	6.5	45.1 ± 2.7%	55.2
CiteGuard+DeepSeek-R1	**95.5**	**87.0**	**71.8**	**15.2**	**68.1 ± 5.8%**	66.7
CiteGuard+Gemini	81.8	43.5	16.7	0.0	34.2 ± 2.7%	40.6
CiteGuard+Kimi-K2	90.9	83.3	41.9	13.0	59.7 ± 2.1%	**68.8**
CiteGuard+Qwen3	81.6	68.8	43.6	10.1	53.1 ± 3.5%	62.5
Human	-	-	-	-	**69.7***	-

Table 2: CiteGuard accuracy in the CiteME benchmark averaged over 5 runs. In this table, "Agree" denotes the percentage of CiteGuard suggested citations that human annotations agree are relevant, and * denotes the number reported by CiteAgent (Press et al., 2024).

Comparison	Δ Acc (%)	$P(\Delta > 0)$	95% CI
vs GPT-4o	+17.7	**0.99**	[8.6, 25.9]
vs Gemini	+31.2	**1.00**	[17.4, 37.7]
vs Qwen3	+18.9	**0.99**	[6.3, 25.1]
vs Kimi-K2	+5.4	0.87	[-3.9, 14.4]

Table 3: Bayesian paired comparisons (vs DeepSeek-R1) among the CiteGuard variants from a single run result.

Method	BioMed(%)	Long Para.(%)	All(%)
CiteAgent	26.6	40.0	31.1
CiteGuard	28.3	46.6	34.4

Table 4: Agent accuracy in the CiteMulti benchmark, averaged over 3 runs using Kimi-K2.

Science domains (i.e., computer vision, natural language processing, algorithms, theory), where each excerpt contains exactly one missing citation. The task is for the LLM agent to suggest an appropriate paper to fill in the missing citation.

3.2 Benchmarking Citation Attribution for Cross-Domain Excerpts

To explore *CiteGuard*'s robustness on other scenarios, we collect 10 long multi-citation paragraphs from Computer Science literature in addition to the CiteME excerpts. For cross-domain robustness, we collect 20 additional excerpts from the biomedical domain, each containing a single missing citation. The test samples are collected from PubMed and manually verified for availability on Semantic Scholar to ensure solution feasibility. Test samples cover various biomedicine domains, including cancer, epidemiology, microbiome, and microbial genomics. The combined dataset is denoted as **CiteMulti**.

3.3 Evaluation Strategies

For all the evaluations, we follow the same hyperparameter settings (e.g., temperature) as CiteAgent (Press et al., 2024).

We evaluate *CiteGuard* on both closed- and open-source models, including non-reasoning (GPT-4o (Hurst et al., 2024), Kimi-K2 (Team et al., 2025), Qwen3 (Yang et al., 2025), Gemini 2.0 Flash (Cloud, 2025)) and reasoning model (DeepSeek-R1 (Guo et al., 2025)), taking the average of 5 run results.

3.4 Difficulty Level Labels

We label the sample with difficulty levels using the following criteria from the results of a single run:

- Easy (22 excerpts): Correct for all models
- Medium (46 excerpts): Correct for more than three out of five models
- Medium-Hard (39 excerpts): Correct for no more than two out of five models
- Hard (23 excerpts): Incorrect for all models

We present some examples of excerpts in the more difficult levels in App. C.

3.5 Human Assessment

To evaluate the quality of alternative citations suggested by CiteGuard, we manually inspect the suggested citations from a single run for all the different settings (see Fig. 16 and Fig. 11 for examples). For each sampled claim, we ask at least two expert annotators with backgrounds in computer science and scientific writing to independently judge whether the list of suggested citations produced by CiteGuard, backed by different LLMs, is an appropriate alternative. We defined an alternative citation as "appropriate" if it provides equivalent or stronger evidence for the scientific claim compared to the original reference. Inter-annotator agreement reports **72.7%**, indicating high consistency among human annotators.

4 Results

4.1 CiteGuard Accurately Grounds Scientific Claims Through Enhanced Actions

Results in Table 2 demonstrate that *CiteGuard* substantially outperforms CiteAgent, improving the accuracy of retrieving the oracle citation by 12.3% on CiteME when both are powered by GPT-4o. When backed by open-source models DeepSeek-R1 and Kimi-K2, *CiteGuard* achieves up to 65.4% accuracy, approaching the 69.7% human performance reported in CiteME (Press et al., 2024).

We also present the Bayesian paired comparison results among CiteGuard variants using per-sample correctness outcomes from a single run in Table 3. As shown in the results, DeepSeek-R1 demonstrates strong posterior evidence of improved accuracy over GPT-4o, Gemini, and Qwen3 (with $P(\Delta > 0) \geq 0.99$ and the 95% credible intervals above zero), while its advantage over Kimi-K2 is smaller and not conclusive ($P(\Delta > 0) = 0.87$ and the 95% credible interval overlaps with zero). Since per-sample predictions are unavailable for the original CiteAgent-reported statistics, Bayesian tests are not applicable.

This improvement is driven by *CiteGuard*'s extended retrieval actions (§ 2.2), which make citation search more flexible and robust. As illustrated in Fig. 1, while CiteAgent relies heavily on the `read` action that assumes reliable PDF access, *CiteGuard* succeeds through introducing two key new actions: (1) `ask_for_more_context` enables the agent to proactively query for additional claim context when the initial snippet is insufficient, and (2) `search_text_snippet` allows searching directly within paper contents.

There are two key advantages of search_text_snippet:

- When relevant information is not explicit in a paper's title or abstract, search_text_snippet can retrieve matches from the main content, whereas earlier search actions rely solely on title and abstract matching
- It does not require PDF access, which is often unavailable or difficult to parse in Semantic Scholar. When PDF access fails, agents may rely only on titles and abstracts and make ill-informed selections. In contrast, search_text_snippet retrieves text directly from paper content and excludes papers without accessible content, leading to more robust retrieval.

This step-by-step reasoning, together with more advanced actions, enables *CiteGuard* to accurately identify the oracle citation where CiteAgent fails, improving the accuracy and robustness of scientific claim grounding, particularly in real-world citation retrieval with complex long-range contexts.

4.2 CiteGuard Effectively Suggests Alternative Citations

Through manual assessment, *CiteGuard* showcases its ability to generate high-quality alternative citations beyond the original reference (Table 2). Concretely, by using aggregated human annotations as a new oracle, Table 2 computes the agreement between *CiteGuard*'s suggested citations and human judgments. Across models, *CiteGuard* achieved substantial alignment with human evaluations, demonstrating its potential to identify relevant alternative literature.

Notably, this ability is **model-agnostic**: both proprietary models like GPT-4o and open-source models like Qwen3 can effectively identify relevant alternatives. Fig. 16 demonstrates *CiteGuard*'s *backward reasoning* ability based on the excerpt. Fig. 11 further shows the *lateral reasoning* capacity of *CiteGuard*, where *CiteGuard* effectively identifies highly related work as the oracle reference suggested. We include some examples and more details about the human assessment process in App. D.

4.3 CiteGuard Can Generalize to Cross-Domain Citation Attribution

Results in Table 4 demonstrate *CiteGuard*'s potential to generalize it's ability to ground scientific claims to cross-domain and long paragraph scenarios. We note that the performance gain of CiteGuard over the CiteAgent baseline on CiteMulti is smaller than that observed on CiteME. This is likely due to domain-specific distribution shifts and differences in citation styles. We discuss the limitations associated with the search engine in the limitations section.

5 Analysis

5.1 Effectiveness of Multi-Run

To evaluate *CiteGuard*'s performance with iterative run ability enabled, we present the result of CiteGuard + Kimi-K2 on the CiteMe benchmark in Table 9. The result demonstrates that CiteGuard with iterative runs consistently improves accuracy, with diminishing yet still meaningful gains as the number of iterations increases. For the excerpt where only one particular citation is appropriate and is selected in prior runs, *CiteGuard* sometimes refuses to suggest a paper after exceeding the number of actions, as the papers from the filtered search result are not relevant.

5.2 Effectiveness of Each Action

To demonstrate the contribution of each newly added action, we perform an ablation study for the two actions **ask_for_more_context** and **search_text_snippet** and present the result below. As shown in Table 7, introducing either action leads to a substantial improvement over the CiteAgent baseline, indicating that both additional context elicitation and external evidence retrieval are effective for citation verification. In particular, the **search_text_snippet** action yields a larger individual gain, suggesting that access to retrieved textual evidence is particularly important. When both actions are enabled, CiteGuard achieves the highest accuracy, indicating that the two actions complement each other.

5.3 Retrieval vs Long-Context

To demonstrate the effect of retrieving only relevant parts of the paper versus providing the full paper text, we run the *CiteGuard*+Kimi-K2 agent, replacing the "find_in_text" action with the "read" action and present the results in Table 8. With the "read" action, the accuracy increased by 3.07%, at the cost of 2× more tokens. The number of tokens can be as large as 4× as shown in App.G.2.

Although reading the full paper content in context can provide some benefits, it is at the cost of significantly more tokens. When using *CiteGuard*, users would need to determine whether to use retrieval or long-context based on the token budget.

5.4 Tradeoffs of Human in the Loop

When we perform the evaluation, ask_for_more_context does not take input from users, but instead returns the paragraph in the source paper that contains the excerpt (with the citation masked). However, for applications where the excerpt is not taken from a paper but provided by a researcher in the process of writing, the user would have to provide some additional context manually if needed.

We perform an ablation study by removing ask_for_more_context action; the performance difference is subtle (2%), but may be important, depending on the use case. Although ask_for_more_context and the full list of actions are provided as default, there is an option to remove some of the actions, making CiteGuard applicable to different use cases.

5.5 Reasoning vs Non-Reasoning Models

Table 2 shows that the difference of open-sourced reasoning (DeepSeek-R1) and non-reasoning model (Kimi-K2) in overall performance; however, its advantage is not conclusive under Bayesian paired testing **(P = 0.87)** as shown in Table 3.

Model	Avg. Input Tokens	Avg. Output Tokens	Avg. Cost / Sample ($)	Platform
GPT-4o	17,931.8	1,705.8	0.12	OpenAI
DeepSeek-R1 (671B / 37B)	**15,004.9**	1,771.4	0.005	DeepSeek
Gemini-2.0-Flash	19,064.9	1,449.4	**0.00**	Google (free tier)
Kimi-K2 (1T / 30B)	15,451.4	**826.7**	0.017	Together AI
Qwen3 (235B / 22B)	14,598.8	936.8	**0.003**	Together AI

Table 5: Average token usage and API cost per sample for each evaluated model. Numbers are averaged over the evaluation set. Model sizes are reported as total parameters / activated parameters where applicable.

# of Iterations	**Accuracy (%)**
1	60.0
2	70.0% (+ 10.0%)
3	75.4% (+ 5.4%)

Table 6: *CiteGuard*+Kimi-K2 accuracy difference on the CiteME benchmark when using iterative runs.

Method	**Accuracy (%)**
CiteAgent	36.2%
CiteGuard (+ ask_for_more_context)	53.1%
CiteGuard (+ search_text_snippet)	57.7%
CiteGuard (+ both)	59.7%

Table 7: Accuracy difference on the CiteME benchmark when using different sets of actions.

Method	**Accuracy (%)**	**Avg # of Tokens**
read	63.1	33,544.68
find_in_text	60.0	15,451.43

Table 8: *CiteGuard*+Kimi-K2 accuracy difference on the CiteME benchmark when using different actions to get information from the paper content.

Method	**Accuracy (%)**
CiteGuard (w/o ask_for_more_context)	57.7%
CiteGuard (w/ ask_for_more_context)	59.7%

Table 9: *CiteGuard*+Kimi-K2 accuracy difference on the CiteME benchmark when using different actions.

	Top 1	**Top 10**
AI2 Paper Finder	38.5	60.0
Ours+Gemini	36.9	44.7
Ours+DeepSeek-R1	**65.4**	**71.5**

Table 10: AI2 Paper Finder (AI2, 2025)'s accuracy (%) on CiteME compared to CiteGuard.

As demonstrated in the example (Fig. 4), a reasoning model tends to question itself (`"But note:...However,..."`) and consider other available actions, while a non-reasoning model would be more confident in its action (`"I can still be confident that..."`). Although the agent backed by both models eventually arrived at different citations, both are deemed correct through human assessment, demonstrating that *CiteGuard* is not dependent on reasoning ability.

5.6 CiteGuard vs Paper Finders

An alternative to finding potential references using *CiteGuard* is to use a paper finder. We run AI2 Paper Finder (AI2, 2025) on CiteME and present the results in Table 10. AI2 Paper Finder searches and ranks the documents, which can result in a long list of papers, while CiteGuard operates in a setting that only produces one suggestion at a time. Therefore, we report AI2 Paper Finder's accuracy by taking the top k-ranked documents, and CiteGuard's top k performance from k separate (non-iterative) runs. We argue that *CiteGuard* matches Paper Finder in terms of accuracy, if not surpassing it. In particular, the top 10 accuracy is 5.4 percentage points below the top 1 accuracy of CiteGuard+DeepSeek-R1, demonstrating that *CiteGuard* is more reliable, which is likely because it incorporates the context of the excerpt.

5.7 Cost Analysis

We provide the average number of tokens and the API cost per sample for each model we evaluated in Table 5. Overall, CiteGuard incurs relatively modest inference cost. The best-performing model, DeepSeek-R1, has an average API cost of approximately $0.005 per citation, which is substantially lower than GPT-4o ($0.12 per citation). These results suggest that CiteGuard's performance improvements are achieved without incurring disproportionate inference cost, supporting its practicality and scalability.

6 Related Work

6.1 Retrieval-Augmented Generation and LLMs for Scientific Research

Retrieval Augmented Generation (RAG) models were first introduced as models that can combine

DeepSeek-R1: ... But note: the excerpt says "the raw image could be fed directly into the reinforcement learning algorithm through a series of convolutions". This is exactly what DQN did. However, let's break it down: - The citation is likely for the method of using convolutions to process raw images in RL, not necessarily the entire RL algorithm...

Kimi-K2: ...I can still be confident that its abstract already declares it is the first deep learning model to successfully learn control policies directly from high-dimensional sensory input, where input is raw pixels...

Figure 4: CiteGuard thought example when backed by reasoning model (DeepSeek-R1) and non-reasoning model (Kimi-K2)

parametric and non-parametric memory (Lewis et al., 2020). Recently, RAG has shown to be a promising direction toward mitigating hallucinations and other challenges in knowledge-intensive tasks for LLMs (Borgeaud et al., 2022; Gao et al., 2023). One application is the use of LLMs/LLM Agents to assist human researchers, such as knowledge discovery, proposing ideas, carrying out experiments, scientific writing, conducting reviews, or even transforming paper into interactive agents (Lu et al., 2024; Yamada et al., 2025; Miao et al., 2025). As part of the effort to mitigate the hallucination issue in LLMs for scientific writing, RAG-aware fine-tuned LLMs for literature summaries have been introduced (Asai et al., 2024a; Wang et al., 2025).

6.2 Citation Suggestion

There have been different approaches to citation recommendation before the era of LLMs, including information retrieval (Färber and Sampath, 2020) and neural networks (Ebesu and Fang, 2017; Yang et al., 2018; Jeong et al., 2020). These methods require re-training and do not account for the rapidly updating paper database. In light of this, LLM agentic workflow (Press et al., 2024) has been proposed to enable access to a real-time paper database.

In this work, we adopted the CiteAgent (Press et al., 2024) framework, where retrieval is performed through tool calls to the Semantic Scholar API, which we treat as a black box for retrieval. This approach would benefit from further improvements in the retrieval pipeline of the API. The framework is built to enable multiple rounds of retrieval and reading, with the choice of action dependent on the agent's own decision following its thought, similar to the ReAct approach (Yao et al., 2023). *CiteGuard* uses RAG to provide better evaluation for LLM-generated literature summaries.

6.3 LLM-as-a-Judge

Evaluation of LLM-generated text has traditionally been carried out by humans. Collecting human annotations is costly and not scalable. To overcome this issue, LLM-as-a-Judge was introduced to automate the evaluation process (Zheng et al., 2023). Due to the improved scalability, LLM-as-a-Judge has been widely used to evaluate LLM-generated scientific writing. For instance, OpenScholar (Asai et al., 2024a) uses LLM-as-a-Judge to filter and refine LLM-synthesized training data. However, LLM-as-a-Judge exhibits bias (Ye et al., 2024; Gu et al., 2024) or sensitivity to prompt (Thakur et al., 2024). Moreover, LLM-as-a-Judge often requires a text snippet of the citation under review, which limits its use case for scenarios where the text snippets used during generation are not available. In this work, we explore expanding LLM-as-a-Judge to include RAG to alleviate biases and provide a more robust evaluation in cases where relevant text snippets are not directly available. A similar idea is Agent-as-a-Judge (Zhuge et al., 2024) targeting the task of automated code generation for AI development.

7 Conclusion and Future Directions

We observe the limitation in using LLM-as-a-Judge for citation attribution of scientific writing and propose *CiteGuard* agent to provide a more faithful citation attribution through retrieval-augmented validation. We show the reliability of CiteGuard in finding correct citations to be approaching human performance, and the alternative citations suggested by *CiteGuard* are deemed relevant by human annotators. We further demonstrate *CiteGuard*'s ability to generalize across domains.

Faithful attribution is critical to the integrity of scientific communication, and the research community is placing increasing emphasis on accurate, verifiable citation attribution. We envision *CiteGuard* as one of the components in future scientific writing and review pipelines, to mitigate hallucinated citations and promote more trustworthy scholarly literature.

Limitations

Currently, the implementation of CiteGuard is based on the Semantic Scholar API, which causes CiteGuard's performance to be limited by the coverage of the database and the ability of the retrieval pipeline of the API. One future direction of CiteGuard is enabling the use of other research literature databases and retrieval pipelines. Although we have shown that CiteGuard agent works well with both open-sourced and closed-sourced, both reasoning and non-reasoning models, we have not yet explored its performance on smaller open-sourced models (e.g., models with less than 1B parameters) due to the limitation of time. We plan to conduct such an analysis and evaluate how much CiteGuard depends on the models' size.

Ethical considerations

Our work aims to promote a more faithful citation attribution for scientific writing, regardless of machine-generated or human-generated. The framework relies on Large Language Models, which may exhibit systemic biases in research communities, such as geographic and linguistic biases. Although our method is model-agnostic, we acknowledge that mitigating these biases is still an open challenge. Future work includes better representation of under-cited or non-English sources. Our framework uses Semantic Scholar, which is an open-access research tool for scientific literature, through its API. We have not used any private or sensitive data. All human annotators (including the authors) participated in a voluntary manner, with their identities kept anonymous during the analysis.

References

AI2. 2025. Introducing ai2 paper finder.

Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D'arcy, and 1 others. 2024a. Openscholar: Synthesizing scientific literature with retrieval-augmented lms. *arXiv preprint arXiv:2411.14199*.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024b. Self-rag: Learning to retrieve, generate, and critique through self-reflection.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, and 1 others. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

Lutz Bornmann and Rüdiger Mutz. 2015. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the association for information science and technology*, 66(11):2215–2222.

Uchendu Eugene Chigbu, Sulaiman Olusegun Atiku, and Cherley C Du Plessis. 2023. The science of literature reviews: Searching, identifying, selecting, and synthesising. *Publications*, 11(1):2.

Google Cloud. 2025. Gemini 2.0 flash.

Travis Ebesu and Yi Fang. 2017. Neural citation network for context-aware citation recommendation. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, pages 1093–1096.

Zhiyuan Fan, Longfei Yun, Ming Yan, Yumeng Wang, Dadi Guo, Brian Mak, James Kwok, and Yi R. Fung. 2025. End-to-end optimization for multimodal retrieval-augmented generation via reward backpropagation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*.

Michael Färber and Ashwath Sampath. 2020. Hybridcite: A hybrid model for context-aware citation recommendation. In *Proceedings of the ACM/IEEE joint conference on digital libraries in 2020*, pages 117–126.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1).

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, and 1 others. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Qi He, Cheng Qian, Xiusi Chen, Bingxiang He, Yi R. Fung, and Heng Ji. 2025. Veri-r1: Toward precise and faithful claim verification via online reinforcement learning. *Preprint*, arXiv:2510.01932.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Chanwoo Jeong, Sion Jang, Eunjeong Park, and Sungchul Choi. 2020. A context-aware citation recommendation model with bert and graph convolutional networks. *Scientometrics*, 124(3):1907–1922.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

Michael Krumdick, Charles Lovering, Varshini Reddy, Seth Ebner, and Chris Tanner. 2025. No free labels: Limitations of llm-as-a-judge without human grounding. *arXiv preprint arXiv:2503.05061*.

Peder Larsen and Markus Von Ins. 2010. The rate of growth in scientific publication and the decline in coverage provided by science citation index. *Scientometrics*, 84(3):575–603.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*.

Jiacheng Miao, Joe R Davis, Jonathan K Pritchard, and James Zou. 2025. Paper2agent: Reimagining research papers as interactive and reliable ai agents. *arXiv preprint arXiv:2509.06917*.

Ori Press, Andreas Hochlehnert, Ameya Prabhu, Vishaal Udandarao, Ofir Press, and Matthias Bethge. 2024. Citeme: Can language models accurately cite scientific claims? *Advances in Neural Information Processing Systems*, 37:7847–7877.

Gerard Salton and Donna Bergmark. 1979. A citation study of computer science literature. *IEEE Transactions on Professional Communication*, (3):146–158.

Nazar Shmatko, Alex Cui, Alex Adam, and Paul Esau. 2025. Gptzero uncovers 50+ hallucinations in iclr 2026.

Hannah Snyder. 2019. Literature review as a research methodology: An overview and guidelines. *Journal of business research*, 104:333–339.

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, and 1 others. 2025. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*.

Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. 2024. Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges. *arXiv preprint arXiv:2406.12624*.

William H Walters and Esther Isabelle Wilder. 2023. Fabrication and errors in the bibliographic citations generated by chatgpt. *Scientific Reports*, 13(1):14045.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. 2024a. SciMON: Scientific inspiration machines optimized for novelty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 279–299, Bangkok, Thailand. Association for Computational Linguistics.

Yidong Wang, Qi Guo, Wenjin Yao, Hongbo Zhang, Xin Zhang, Zhen Wu, Meishan Zhang, Xinyu Dai, Qingsong Wen, Wei Ye, and 1 others. 2024b. Autosurvey: Large language models can automatically write surveys. *Advances in neural information processing systems*, 37:115119–115145.

Yubo Wang, Xueguang Ma, Ping Nie, Huaye Zeng, Zhiheng Lyu, Yuxuan Zhang, Benjamin Schneider, Yi Lu, Xiang Yue, and Wenhu Chen. 2025. Scholarcopilot: Training large language models for academic writing with accurate citations. *arXiv preprint arXiv:2504.00824*.

Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. 2025. The ai scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv preprint arXiv:2504.08066*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Libin Yang, Yu Zheng, Xiaoyan Cai, Hang Dai, Dejun Mu, Lantian Guo, and Tao Dai. 2018. A lstm based model for personalized context-aware citation recommendation. *IEEE access*, 6:59618–59627.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, and 1 others. 2024. Justice or prejudice? quantifying biases in llm-as-a-judge. *arXiv preprint arXiv:2410.02736*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Mingchen Zhuge, Changsheng Zhao, Dylan Ashley, Wenyi Wang, Dmitrii Khizbullin, Yunyang Xiong, Zechun Liu, Ernie Chang, Raghuraman Krishnamoorthi, Yuandong Tian, and 1 others. 2024. Agent-as-a-judge: Evaluate agents with agents. *arXiv preprint arXiv:2410.10934*.

A CiteGuard

A.1 Prompts

The system prompt (Fig. 5) and examples provided in the prompt for each newly added actions (Fig. 6, Fig. 7 and Fig. 8) are presented below.

B Examples of CiteGuard Short Trajectories

We evaluate the risk of contamination which means the models are aware of the citation beforehand and does not use search tools to accomplish the task. We manually select some successful short trajectories which are more likely to be an indication of contamination and put the examples in Fig. 9 and 10. Although these successful trajectories are short, we have not found any instances where the agent knows the ground-truth citation in advance and directly searches for the target citation. Instead, in these trajectories, both agents compose a generic search query and identify the appropriate references from the list of search results.

C Difficulty Level Labels

- Easy: Several studies demonstrate the fragility of convolutional networks on simple corruptions. For example, [CITATION] apply impulse noise to break Google's Cloud Vision API. (Ground-Truth: Google's cloud vision api is not robust to noise)
- Medium: To address this, [CITATION] introduced Adversarial Filtering (AF). An overview is shown in Figure 2. The key idea is to produce a dataset D which is adversarial for any arbitrary split of (D_train), (D_test). (Ground-Truth: Swag: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference)
- Medium-Hard: Even if we assume fixed filters using a combination of the above, our probabilistic formulation still allows learning the parameters of the GSM experts from data as outlined below. Consequently, we do not need to tune the trade-off weights between the brightness and gradient constancy terms by hand as in [CITATION]. (Ground-Truth: High Accuracy Optical Flow Estimation Based on a Theory for Warping)
- Hard: RCA [CITATION] is intermediate between PCA and LDA in its use of labeled data. Specifically, RCA makes use of so-called "chunklet" information, or subclass membership assignments. (Ground-Truth: Adjustment learning and relevant component analysis)

D Human Assessment On CiteGuard Alternative Citations

Examples in Fig. 1, 11, and 16, suggest that CiteGuard's extended retrieval actions and strategies not only improve the accuracy of original citation retrieval, but also expand the searching capacity to identify functionally equivalent references, supporting richer scholarly grounding with enhanced accuracy and robustness. Importantly, our manual analysis (Table 2 and Figures 11 & 16) reveals that CiteGuard is capable of both **lateral reasoning** (Fig. 11) and **backward reasoning** (Fig. 16), behaviors that traditional citation retrieval systems typically lack.

- **Backward Reasoning:** While focusing on more recent publications, CiteGuard is capable of identifying citations of previous years written by the same author (Fig. 16).
- **Lateral Reasoning:** CiteGuard suggests peer or related work along with its identification of best-match citations (Fig. 11), providing effective citation finding and alternative suggestions.

E LLM-As-A-Judge Failure

Evaluation Prompt. For the evaluation of OpenScholar citation attribution, we guide the LLM judge through the prompt in Fig. 12.

Failure Example We show how LLM judge can fail in its evaluation of assessment as a result of missing terminology nuances (Fig.13).

You are given an excerpt from a paper, where a citation was deleted. I'm trying to find the citation (ignore the word [CITATION], that's just where the citation was deleted from. You will be asked to help me find the paper from which the citation was deleted. You are equipped with the following tools that will help you in your task: you can search, you can select to find a keyword in a paper from the search results, or you can select a paper as your final answer.
<FORMAT_INSTRUCTIONS>
Your responses have to include one of the actions above. Before you take any action, provide your thoughts for doing so. Do not include anything other than your thoughts and an action in your responses. You must include exactly one action in your responses.
Keep in mind that you can only find keyword in papers or select papers after you search. You can always search, and then search again. You can also find different keywords in a few papers consecutively, without searching again (as long as the papers appeared in your last search).
If your search does not return any relevant results, please try the following: 1. Adjust your query to focus on individual parts of the claim separately, rather than the entire sentence. It is likely that the citation supports only the immediate preceding concept, not the full claim. Use simpler, more general search queries with fewer words (around 3). 2. Avoid overly specific or plural terms—use base forms of key concepts to improve match quality. For example, "We apply contrastive learning to improve the representation quality of a ResNet-based encoder [CITATION].", the citation is most likely for ResNet instead of covering both contrastive learning and ResNet. 3. Change to search by text snippets. You should respond with the following, and use the paper title to do subsequent search.

Figure 5: CiteGuard System Prompt

F LLM Generation Failure

By examining LLM-generated outputs, we also observe failures due to their lack of important elements. For example, Fig. 14 illustrates an LLM generation failure as a result of missing alternative citations.

G Examples of CiteGuard

G.1 Suggestion On Alternatives.

CiteGuard is capable of suggesting meaningful alternatives (Fig.11, and Fig.16). An example of a case where alternative citations are not appropriate is as follows.

- Zephyr-7B-Beta [CITATION] is an instruction-tuned version of Mistral-7B. (Ground-Truth: Zephyr: Direct Distillation of LM Alignment)

G.2 Retrieval vs Long-Context.

We present an example of the CiteGuard+GPT-4o agent when using the "read" action instead of the "find_in_text" action in Fig. 15, where the number of tokens can be as large as 4x. This is due to the additional tokens required when reading multiple full papers in context.

H Human Annotators

All human annotators are graduate students pursuing master's or doctoral computer science degrees at universities where English is the primary language of instruction.

The annotators evaluated suggested citations along two axes:

- **Relevance:** Whether the cited paper genuinely supports the claim.
- **Sufficiency:** Whether the suggested citation can reasonably replace the original in scholarly writing.

Each human annotator was informed that the data collected would be used for a paper submission. The instruction given to the human annotators is as follows:

Please review each excerpt below and:

1. Select all papers that would be suitable for use as citations in the given excerpt context
2. If none of the papers are suitable, please choose "None of the above"

The excerpt is from paper title "Learning Deep Features for Discriminative Localization": In this section, we evaluate the localization ability of CAM when trained on the ILSVRC 2014 benchmark dataset [CITATION]
You would respond with: { "reason": "The excerpt is not sufficient to determine whether the citation refers to the ILSVRC 2014 benchmark dataset or the CAM's localization ability. Let's ask for more context to clarify the scope of the citation.", "action": { "name": "ask_for_more_context", "query": "In this section, we evaluate the localization ability of CAM when trained on the ILSVRC 2014 benchmark dataset [CITATION]" "paper_title": "Learning Deep Features for Discriminative Localization" } }
You will then be given the following input: 3. Weakly-supervised Object Localization In this section, we evaluate the localization ability of CAM when trained on the ILSVRC 2014 benchmark dataset [20]. We first describe the experimental setup and the various CNNs used in Sec. 3.1. Then, in Sec. 3.2 we verify that our technique does not adversely impact the classification performance when learning to localize and provide detailed results on weakly-supervised object localization.
You would respond with: { "reason": "With the additional context, it is now clear that the citation refers specifically to the ILSVRC 2014 benchmark dataset. The cited paper seems to be the ImageNet dataset paper. We'll do a search to make sure.", "action": {...} }

Figure 6: CiteGuard ask_for_more_context Prompt

You would then reply: { "reason": "Paper cbb19236820a96038d000dc629225d36e0b6294a seems to directly mention the ILSVRC 2014 benchmark in the abstract. Though we can use ILSVRC 2014 as the query, but fewer words is better, so we only use ILSVRC as the query.", "action": { "name": "find_in_text", "paper_id": "cbb19236820a96038d000dc629225d36e0b6294a" "query": "ILSVRC" }}
You would then be given the following input:
Based on this work, we attended the competition of ILSVRC 2014 [26]

Figure 7: CiteGuard find_in_text Prompt

You should respond with the following, and use the paper title to do subsequent search: { "reason": "Due to the unsuccessful paper search, let's search by text snippets instead", "action": { "name": "search_text_snippet", "query": "ILSVRC 2014" } } You will then be given the following input: - Title: Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition Section: Abstract Snippet: In ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2014, our methods rank #2 in object detection and #3 in image classification among all 38 teams.

Figure 8: CiteGuard search_text_snippet Prompt

I Use of AI Assistants

The AI assistant (i.e., Grammarly) is used for the writing of this manuscript. All content was critically reviewed and revised by human authors to ensure scientific accuracy and originality.

system: ...

user: You are now given an excerpt. Find me the paper cited in the excerpt, using the tools described above. Please make sure that the paper you select really corresponds to the excerpt: there will be details mentioned in the excerpt that should appear in the paper. If you read an abstract and it seems like it could be the paper we're looking for, read the paper to make sure. Also: sometimes you'll read a paper that cites the paper we're looking for. In such cases, please go to the references in order to find the full name of the paper we're looking for, and search for it, and then select it.
In the spirit of transductive bandits [CITATION] we consider a more general setting where answers are sets of arms. The set of actions and the set of answers can be different.

assistant: {"reason": "The excerpt mentions 'transductive bandits' as the key concept that the citation refers to. Since this is a specific term, I'll start by searching for papers on transductive bandits to find the foundational work.", "action": {
"name": "search_relevance",
"query": "transductive bandits"
}}

user: ...

assistant: {"reason": "The excerpt introduces 'transductive bandits' as an existing concept and extends it to a setting where answers are sets of arms. Among the search results, Paper ID 4f0d485cbcde840533f23f0c8b0f3fa1ca2d74df directly introduces the transductive linear bandit problem, explicitly defining the scenario where measurement vectors (actions) and items (answers) can be different sets. This matches the excerpt's description of distinct action/answer sets. Other papers either apply transductive learning in different domains (e.g., text classification, video segmentation) or focus on variants like federated bandits without defining the core concept.", "action": { "name": "select", "paper_id": "4f0d485cbcde840533f23f0c8b0f3fa1ca2d74df" }}

Figure 9: CiteGuard+Deepseek Short trajectory (history length: 5)

system: ...

user: You are now given an excerpt. Find me the paper cited in the excerpt, using the tools described above. Please make sure that the paper you select really corresponds to the excerpt: there will be details mentioned in the excerpt that should appear in the paper. If you read an abstract and it seems like it could be the paper we're looking for, read the paper to make sure. Also: sometimes you'll read a paper that cites the paper we're looking for. In such cases, please go to the references in order to find the full name of the paper we're looking for, and search for it, and then select it.
Our evaluation follows the protocol of [CITATION], a recent inpainting model that introduces a specialized architecture relying on Fast Fourier Convolutions [8]

assistant: Okay, now I need to locate the paper that (a) is specifically an image-inpainting model and (b) centres its method on Fast Fourier Convolutions. Because the excerpt does not specify the title, I'll start by issuing a highly-targeted search that includes both key phrases.{"reason": "The excerpt identifies the missing citation as recent inpainting model that introduces a specialised architecture relying on Fast Fourier Convolutions [8]. The query must combine both aspects inpainting task and Fast Fourier Convolutions to find the exact paper.", "action": {"name": "search_relevance", "query": "Fast Fourier Convolution inpainting model"}}

user: ...

assistant: {"reason": "The excerpt calls for the recent inpainting model whose key novelty is Fast Fourier Convolution, and abstract fdf7012ebe9d4c4d2d93004613e7a19f69a83a93 directly fits that description: it introduces a network that uses fast Fourier convolutions (FFCs) to obtain image-wide receptive fields for large-hole inpainting.", "action": { "name": "select", "paper_id": "fdf7012ebe9d4c4d2d93004613e7a19f69a83a93" }}

Figure 10: CiteGuard+Kimi-K2 Short trajectory (history length: 5)

Excerpt: For a comprehensive evaluation, we test on 19 datasets spanning a wide range of object, scene and fine-grained categories: ImageNet [18], StanfordCars [43], UCF101 [68], Caltech101 [25], Caltech256 [32], Flowers102 [56], OxfordPets [CITATION], Food101 [7], SUN397 [75], DTD [14], EuroSAT [37], FGVCAircraft [51], Country211 [61], CIFAR-10 [44], CIFAR-100 [44], Birdsnap [5], CUB [72], ImageNet-Sketch [73] and ImageNet-R [38]."

Oracle Reference: Cats and dogs (2012)
CiteGuard Suggestion: The truth about cats and dogs (2011)

Figure 11: Example of CiteGuard Suggested Alternative Citations

As an Attribution Validator, your task is to verify whether a given reference can support the given claim. A claim can be either a plain sentence or a question followed by its answer. Specifically, your response should clearly indicate the relationship: Attributable, Contradictory or Extrapolatory. A contradictory error occurs when you can infer that the answer contradicts the fact presented in the context, while an extrapolatory error means that you cannot infer the correctness of the answer based on the information provided in the context.

Claim: claim
Reference: reference

Figure 12: OpenScholar citation attribution evaluation prompt to LLM

Claim: This technique of lower bounding mutual information is known as **Variational Information Maximization** [CITATION].

Reference: The IM Algorithm: A **variational approach to Information Maximization**. . . .

LLM judgement: Extrapolatory. Cannot infer the correctness of the answer based on the information provided in the context.

Figure 13: LLM mistakenly judges a correct citation correct due to the slight difference in terminology

However, LLMs . . . may lack **comparative analysis**, organizational structure, and key elements (Li et al. 2024). To address these challenges, . . . , **ChatCite**, . . . (Li et al. 2024).

Figure 14: Example of an issue in a LLM-generated text: missing alternative citations (multiple papers other than ChatCite also address comparative analysis)

Excerpt: A second consideration is how to treat the image itself: the raw image could be fed directly into the reinforcement learning algorithm through a series of convolutions[CITATION].

CiteGuard-Read:
Total Input Tokens(75900)
Select: Reward learning from human preferences and demonstrations in Atari

CiteGuard-Find-In-Text:
Total Input Tokens(19182)
Select: Fully Convolutional Network with Multi-Step Reinforcement Learning for Image Processing

Figure 15: CiteGuard example when using "read" vs "find_in_text"

Excerpt: We learn π using behavioral cloning [CITATION], which optimizes π by minimizing the negative log-likelihood of actions given the images and language instructions.

Oracle reference:
Alvinn: An autonomous land vehicle in a neural network (1988)
CiteGuard suggestion:
A Framework for Behavioural Cloning (2001)

Figure 16: Alternative citation suggested by CiteGuard, both are relevant according to human annotations.