Poster: GHOSTCITE: A Large-Scale Analysis of Citation Validity in the Age of Large Language Models

Zuyao Xu[†*], Yuqi Qiu[†*], Lu Sun[†*], Fasheng Miao[‡*], Fubin Wu[†], Xinyi Wang[†], Xiang Li[†✉], Haozhe Lu[†], Zhengze Zhang[†], Yuxin Hu[†], Jialu Li[†], Jin Luo[†], Feng Zhang[‡], Rui Luo[†] Xinran Liu[†], Yingxian Li[†], Jiaji Liu[†]

[†]Nankai University [‡]Tsinghua University **Equal contribution* ✉ *Corresponding author.*

***Abstract*—Citations provide the basis for trusting scientific claims; when they are invalid or fabricated, this trust collapses. With Large Language Models (LLMs) increasingly used for academic writing, their tendency to fabricate citations ("ghost citations") poses a systemic threat to citation validity. To quantify this threat, we develop CITEVERIFIER, an open-source framework for large-scale citation verification. We conduct a comprehensive study consisting of an LLM benchmark, an archival analysis of over 2.2 million citations from top-tier AI and Security venues (2020–2025), and a behavioral user survey. Our findings reveal an accelerating crisis where unreliable AI tools, combined with inadequate human verification, enable fabricated citations to contaminate the scientific record.**

1. Introduction

Citations serve as a core trust mechanism that justifies scientific claims and reliably grounds future research. The rapid advent of AI-assisted academic writing has introduced a novel and systemic threat to this essential scholarly trust infrastructure. Researchers increasingly leverage Large Language Models (LLMs) to synthesize literature reviews, draft related work sections, or compile reference lists. However, driven by their autoregressive generative nature, these AI systems frequently synthesize references by combining statistically correlated tokens into citations that appear perfectly authentic despite being entirely fabricated. This dangerous phenomenon, commonly known as "ghost citations," actively misdirects research efforts and creates a deceptive illusion of evidential support.

Despite growing awareness and explicit policy warnings from major academic publishers and conference organizers, the scientific community remains vulnerable. Three critical practical gaps remain unresolved: (1) Scalable automated detection is inherently difficult due to heterogeneous citation formats and incomplete bibliographic databases; (2) We currently lack large-scale empirical measurements determining how prevalent these ghost citations actually are within the published scientific record; and (3) We lack a clear explanation for why existing human verification practices systematically fail to catch these errors during peer review. This poster directly addresses these outstanding gaps by introducing a robust automated verification framework and presenting comprehensive empirical findings on the prevalence, temporal surge, and underlying human factors associated with ghost citations.

2. Methodology: CITEVERIFIER

As shown in Figure 1, to systematically analyze ghost citations at scale, we developed CITEVERIFIER [1], an automated, open-source verification framework. Engineered to process millions of records asynchronously, it overcomes formatting heterogeneity, database inconsistencies, and API limits via a highly concurrent three-stage pipeline:

- **Hybrid Reference Parsing:** To handle format diversity, extraction noise, and OCR errors in raw PDFs, we employ a hybrid approach. We use GROBID as the primary parser, with a Qwen-based LLM serving as a reliable fallback to extract structured metadata into a standardized JSON schema.
- **Cascaded Multi-Source Verification:** To maximize coverage scalably, we use a sequential search strategy. The system first queries a local SQLite cache and DBLP database. For uncached entries, it checks academic databases (e.g., Google Scholar via proxy APIs), falling back to general web searches and LLM-assisted query correction if needed.
- **Similarity-Based Classification:** Importantly, our goal is to identify fully fabricated references rather than benign metadata errors (e.g., minor typos in titles or author names). We classify citation validity using the Levenshtein distance between normalized input titles and retrieved candidates. An empirically calibrated threshold ($\theta = 0.9$) tolerates minor format variations while strictly flagging nonexistent ghost works.

Pipeline Validation. To validate the reliability of CITEVERIFIER, we manually checked two random samples of 400 valid and 400 invalid citations. Valid accuracy was 100%, and invalid accuracy was 98% (392/400); the remaining 8 cases were non-paper sources (4), out-of-domain works (1), or poorly indexed items (3), confirming the high precision of our pipeline.

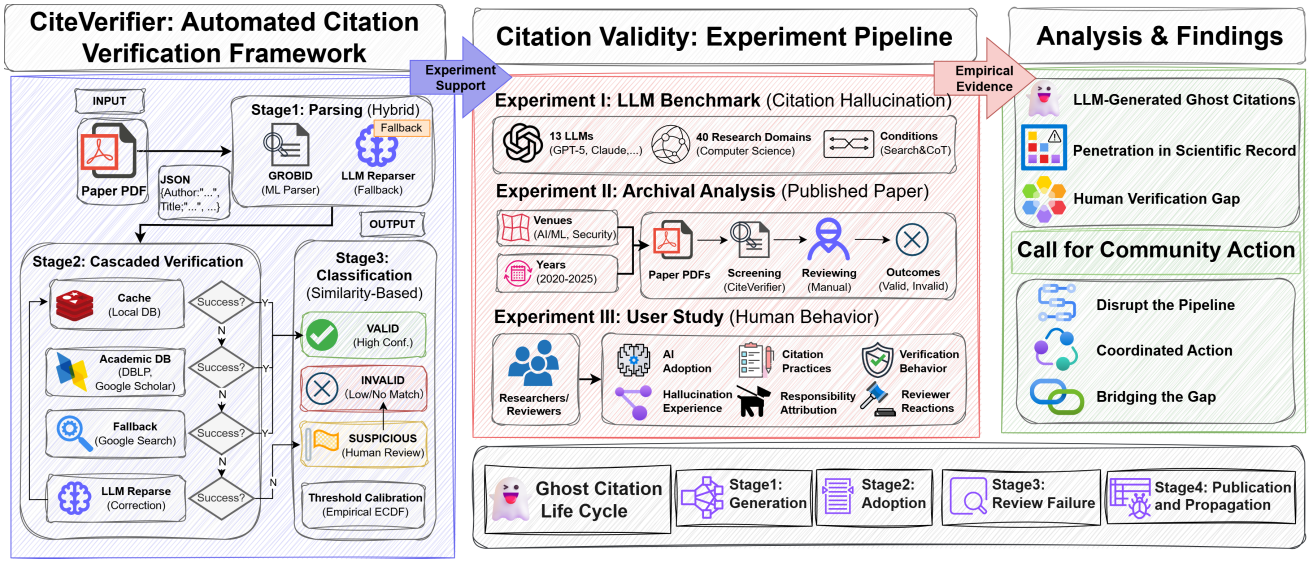


Figure 1: The framework of CITEVERIFIER and experimental pipeline.

3. Experiments and Preliminary Results

Using CITEVERIFIER, we conducted a comprehensive multi-dimensional analysis of the ghost citations.

LLM Benchmark. We evaluated 13 state-of-the-art LLMs (including GPT-5 and DeepSeek) across 40 computer science research domains, generating 375,440 citations.

- **Hallucination Rates:** All models hallucinated citations, with rates ranging from 14.23% (DeepSeek) to 94.93% (Hunyuan).
- **Domain Sensitivity:** Hallucination rates varied widely across domains, with a 51.39 percentage point spread between the best and worst-performing domains.
- **Validation Failure:** When prompted to evaluate citation validity, LLMs achieved only 38% average accuracy, worse than random guessing.

Archival Analysis. We collected 56,381 published papers from eight top-tier AI/ML and Security venues (NeurIPS, ICML, IJCAI, AAAI, IEEE S&P, USENIX Security, CCS, NDSS) from 2020 to 2025. We extracted 2,199,409 citations and processed them through our framework, followed by rigorous manual verification of flagged citations by a 16-person team.

- **Prevalence:** 1.07% of papers (604 papers) contained at least one invalid or fabricated citation.
- **Temporal Surge:** We observed an 80.9% increase in the invalid citation rate in 2025 compared to the 2020–2024 average, correlating with the rise of autonomous AI agent workflows.
- **Error Propagation:** We identified repeated invalid citations (e.g., one erroneous title propagating to 16 independent papers), demonstrating how mistakes diffuse across the literature.

User Study. We surveyed 94 active researchers and reviewers to understand why fabricated citations reach publication.

- **Risky Behaviors:** While 86.7% of AI users claim to "always verify" AI-generated citations, 41.5% of respondents copy-paste BibTeX without checking, and 17.3% cite AI-suggested papers without reading them.
- **Reviewer Blindspots:** 76.7% of reviewers do not thoroughly check references, and 80.0% never suspect fake citations in submissions. Furthermore, 74.5% of researchers view peer review as ineffective at catching metadata errors.

4. Conclusion and Mitigation

Our findings highlight a self-reinforcing "pollution pipeline" in which unreliable LLM-generated content, researcher cognitive offloading, and trust-based review norms allow fabricated citations to persist. We strongly recommend that conferences deploy automated reference checking at submission (supported by 70.2% of surveyed researchers), and that AI developers ground outputs in verified retrieval sources rather than parametric memory.

Acknowledgment. Authors from Nankai University were supported by the National Natural Science Foundation of China (No. 62502236, No. U25B2025), the Natural Science Foundation of Tianjin (No. 24JCQNJC02070), and the Open Project of National Engineering Laboratory for Technology of Internet Domain Name (No. KF202516). Authors from Tsinghua University were supported by the National Key Research and Development Program of China under Grant No.2020YFE0200500.

References

[1] Z. Xu, Y. Qiu, L. Sun, F. Miao, F. Wu, X. Wang, X. Li, H. Lu, Z. Zhang, Y. Hu, J. Li, J. Luo, F. Zhang, R. Luo, X. Liu, Y. Li, and J. Liu, "Ghostcite: A large-scale analysis of citation validity in the age of large language models," 2026. [Online]. Available: https://arxiv.org/abs/2602.06718

Poster: GhostCite: A Large-Scale Analysis of Citation Validity in the Age of Large Language Models

Zuyao Xu [1*] Yuqi Qiu [1*] Lu Sun [1*] Fasheng Miao [2*] Fubin Wu [1] Xinyi Wang [1]
Xiang Li [1✉] Haozhe Lu [1] Zhengze Zhang [1] Yuxin Hu [1] Jialu Li [1] Jin Luo [1]
Feng Zhang [1] Rui Luo [1] Xinran Liu [1] Yingxian Li [1] Jiaji Liu [1]

[1] Nankai University [2] Tsinghua University
*Equal contribution. ✉ Corresponding author.

The Framework of CiteVerifier [1] and Experiment Design

- **Parsing Stage.** Extract metadata from PDF citation strings with GROBID. If parsing fails, Qwen3-Flash re-extracts key fields such as the title.
- **Verification Stage.** Query the local cache, DBLP, and Google Scholar in sequence. If all fail, the system re-parses the citation and retries.
- **Classification Stage.** Compare the input title with the best retrieved title using normalized Levenshtein similarity. Citations below 0.9 are labeled as ghost citations.
- **Reliability Check.** We manually reviewed 400 ”valid” citations (95% confidence, 5% margin) and found no additional invalid ones.

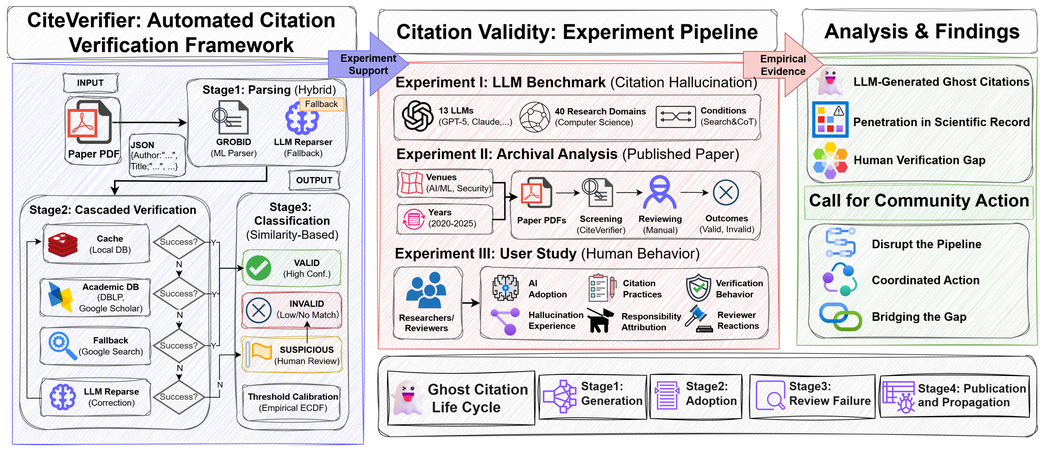


Figure 1. The framework of CiteVerifier and experimental pipeline.

We deploy CiteVerifier in three complementary experiments to investigate ghost citations:

- **LLM Benchmark.** Evaluate 13 LLMs across 40 CS domains (about 375K generated citations) to quantify hallucination rates under varied prompting conditions.
- **Archival Analysis.** Analyze 56,381 papers from top AI and Security venues (2020-2025), combining automated screening with manual verification to measure real-world prevalence.
- **User Study.** Survey active researchers on citation practices and verification behaviors to understand human factors enabling ghost citations to persist.

Measuring the Unreliability: LLM Benchmark

- **High Hallucination Rates.** All 13 LLMs show substantial citation hallucination, from 14.23% to 94.93%.
- **Domain Sensitivity.** Performance varies widely across domains. Some models do well in one area but fail in others.
- **Poor Validation Performance.** LLMs are unreliable at verifying citation validity, achieving only 38% accuracy.

Table 1. LLM performance on citation generation.

Model	Valid JSON (%)	Citations	Halluc. (%)	95% CI
DeepSeek	97.67	27,973	14.23	±1.65
GLM-4.5	97.22	27,835	21.25	±1.94
Claude 4	100.00	28,800	21.84	±1.93
Qwen-3	99.38	28,546	23.52	±1.99
Kimi	89.09	24,955	41.86	±2.44
Llama 4	97.33	27,925	45.84	±2.36
GPT-5	97.90	28,366	50.92	±2.36
Gemini	95.80	27,102	59.47	±2.34
ERNIE	95.06	27,332	71.90	±2.15
Seed	48.75	11,090	78.64	±2.74
Grok 4	99.03	28,627	79.98	±1.88
Phi-4	93.35	27,164	87.47	±1.60
Hunyuan	62.90	16,094	94.93	±1.29

Halluc.: Hallucination rate, the percentage of citations verified as invalid.
95% CI: Confidence interval margin (±) for the hallucination rate.

Table 2. Citation validity judgment accuracy.

Model	Valid Recall	Invalid Precision	Accuracy
ERNIE	6/50 (12%)	50/50 (100%)	56/100 (56%)
Hunyuan	19/50 (38%)	29/50 (58%)	48/100 (48%)
Seed	31/50 (62%)	16/50 (32%)	47/100 (47%)
Phi-4	12/50 (24%)	29/50 (58%)	41/100 (41%)
Kimi-K2	14/50 (28%)	26/50 (52%)	40/100 (40%)
GPT-5	8/50 (16%)	28/50 (56%)	36/100 (36%)
Llama-4	19/50 (38%)	16/50 (32%)	35/100 (35%)
DeepSeek	16/50 (32%)	19/50 (38%)	35/100 (35%)
Qwen-3	13/50 (26%)	21/50 (42%)	34/100 (34%)
Gemini	20/50 (40%)	12/50 (24%)	32/100 (32%)
Grok-4	14/50 (28%)	18/50 (36%)	32/100 (32%)
GLM-4.5	19/50 (38%)	14/50 (28%)	33/100 (33%)
Claude-4	12/50 (24%)	15/50 (30%)	27/100 (27%)
Average	15/50 (30%)	23/50 (46%)	38/100 (38%)

LLMs evaluated on manually verified ground truth from archival analysis (50 valid + 50 invalid entries).

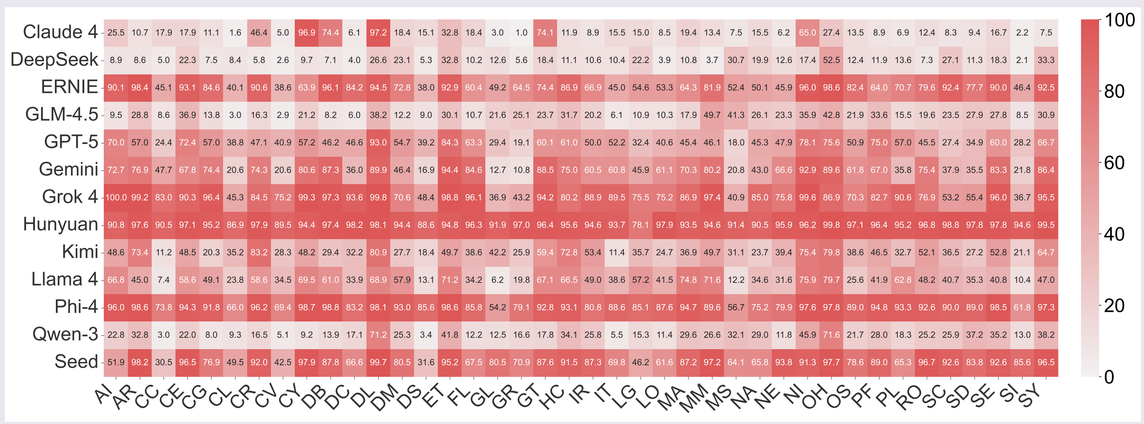


Figure 2. Model-domain hallucination rate heatmap.

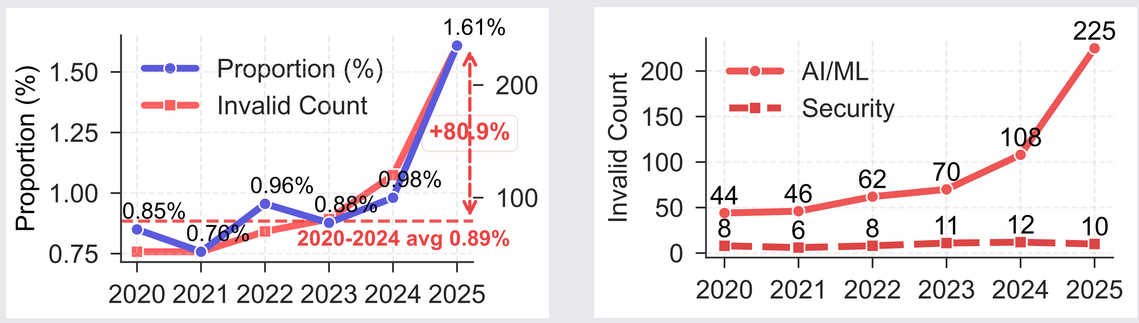
Hunting for Ghosts: Archival Analysis

- **Ghost citations have entered top venues.** Across 2.2M citations from 56,381 papers (2020-2025) across eight leading AI and security venues, we found 604 papers with invalid references after manual review of 2,530 CiteVerifier flagged citations.
- **Two failure types matter.** We found 603 ghost citations, which point to unverifiable works, and 135 metadata errors, which point to real works with wrong citation fields. At the paper level, 486 papers contain ghost citations and 133 contain metadata errors.
- **The share of papers with invalid citations rose in 2025.** From 2020 to 2024, the proportion stayed between 0.76% and 0.98%, then increased to 1.61% in 2025, with up to nine ghost citations in a single manuscript.
- **Errors can spread across the community.** By tracing citation chains, we found that a single metadata error from OpenReview citation export was replicated in at least 16 independent conference papers.

Table 3. Invalid citations by conference.

Conf.	Error	Ghost	Invalid	Papers	Rate	Conf.	Error	Ghost	Invalid	Papers	Rate
NeurIPS	59	332	391	308	1.51%	NDSS	4	19	23	18	2.56%
ICML	22	103	125	104	0.93%	CCS	11	9	20	20	1.14%
AAAI	21	85	106	86	0.62%	USENIX	6	6	12	11	0.57%
IJCAI	11	42	53	51	0.96%	S&P	1	7	8	6	0.56%
Total	135	603	738	604	1.07%						

Invalid: Error + Ghost citations.
Rate: Percentage of papers with invalid citations.
Conf.: Conference acronym.



(a) Overall Time Trend. (b) Time Trend by Venue.

Figure 3. Time trends of papers with invalid citations.

Analyzing Human Factors: Survey Findings

- We collected 94 valid responses from researchers across AI and security. Among them, 82 had published papers and 32 had reviewer experience.
- **Widespread AI Adoption.** 87.2% of respondents use AI tools primarily for text polishing and integrate AI-generated content, including citations, into their writing workflows.
- **Verification Gap.** Despite reporting that they verify references, 41.5% of respondents copy and paste BibTeX without checking, and 76.7% of reviewers do not thoroughly check reference lists.
- **Ineffective Peer Review.** 74.5% view peer review as ineffective at catching citation errors, and 80.0% of reviewers rarely suspect fake citations.
- **High Support for Automation.** 70.2% strongly support deploying automated DOI and reference checks in submission systems.

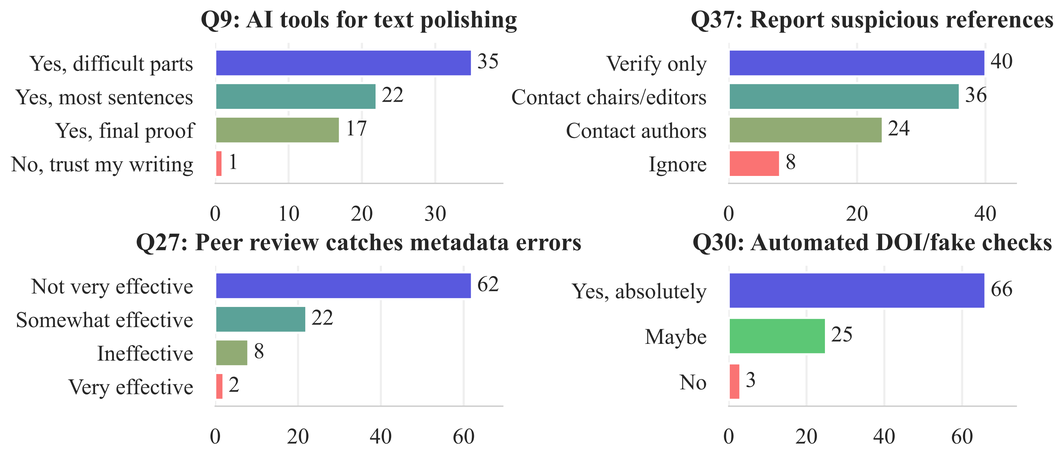


Figure 4. Key questions on AI adoption, reporting behavior, peer review efficacy, and support for automated checks.

References

[1] Zuyao Xu, Yuqi Qiu, Lu Sun, FaSheng Miao, Fubin Wu, Xinyi Wang, Xiang Li, Haozhe Lu, ZhengZe Zhang, Yuxin Hu, et al. Ghostcite: A large-scale analysis of citation validity in the age of large language models. *arXiv preprint arXiv:2602.06718*, 2026.

lixiang@nankai.edu.cn