SynapseCite: Detecting Fabricated Citations in LLM-Generated Literature Reviews

Priya Nandakumar[1] Tomasz Wróbel[2] Aline Dubois[3]

[1]Synapse Labs, NLP Research Group [2]Nexa AI Research [3]Meridian University, Department of Computer Science

p.nandakumar@synapse-labs.example t.wrobel@nexa-ai.example

Proceedings of the 4th Workshop on Scholarly Document Processing and Trustworthy NLP (SDP-Trust), co-located with ACL 2026

Abstract. Literature-review sections drafted by large language models routinely cite papers that do not exist, or attribute real claims to the wrong paper. We present SynapseCite, a lightweight verifier that checks each in-text citation in an LLM-generated review against a retrieved evidence passage from the cited work, rather than trusting the model's own claim of support. SynapseCite combines a bibliographic existence check, a passage retriever over a 2.1M-paper index, and a compact claim–evidence entailment classifier trained on a new corpus of 6,400 citation–claim pairs spanning fabricated, misattributed, and faithful cases. Applied to literature reviews generated by three instruction-tuned LLMs across four research domains, SynapseCite flags fabricated and misattributed citations at 91.4 F1, outperforming a retrieval-only baseline by 11.2 points and a prompted-LLM judge by 6.7 points, while running at a fraction of the judge's inference cost. We release the annotated corpus and discuss where the entailment classifier still fails.

1 Introduction

Ask an instruction-tuned language model to draft the related-work section of a paper and it will produce fluent, well-organized prose complete with in-text citations – and a meaningful fraction of those citations will be wrong. Three failure modes recur: the cited paper does not exist at all (*fabrication*); the paper exists but does not support the claim attributed to it (*misattribution*); and the paper exists and the claim is broadly in its territory, but the specific number, method name, or finding quoted is incorrect (*distortion*). All three are difficult for a human reviewer to catch quickly, because the surrounding sentence reads as confidently as a faithful one.

This matters beyond academic writing assistants. Systems that summarize a research area, generate grant background sections, or produce patent-prior-art surveys inherit the same failure modes, and the downstream reader – a program officer, a patent examiner, an undergraduate doing a literature search – is exactly the audience least equipped to spot a fabricated citation on sight.

Existing defenses fall into two camps. Retrieval-only checks confirm that a cited paper *exists* in a bibliographic database, which catches fabrication but not misattribution or distortion (Oyelaran and Castellano, 2022). Prompted-LLM judges ask a second model whether a claim is supported by an abstract, which is more flexible but expensive at scale and prone to the same overconfidence problems the underlying model class exhibits (Nakamura and Bergstrom, 2021). Neither line of work directly targets the citation-verification setting, where the object to check is a (claim, cited-paper) pair rather than a free-standing factual statement.

We present SynapseCite, a citation verifier purpose-built for LLM-generated literature reviews. Given a sentence and its in-text citation, SynapseCite (i) confirms the cited work exists in a 2.1M-paper index, (ii) retrieves the passage of that paper most relevant to the claim, and (iii) runs a compact entailment classifier over the (claim, passage) pair to decide whether the citation is faithful, misattributed,

or distorted. We train and evaluate the entailment stage on a new corpus, CiteCheck-6K, of 6,400 claim–citation pairs spanning all three error types plus faithful controls, collected from literature reviews generated by three instruction-tuned models across four research domains.

Our contributions are:

- CiteCheck-6K, an annotated corpus of claim–citation pairs labeled fabricated, misattributed, distorted, or faithful, drawn from real model outputs rather than synthetically corrupted citations;
- SynapseCite, a three-stage verifier that outperforms a retrieval-only baseline by 11.2 F1 and a prompted-LLM judge by 6.7 F1, at roughly one-eighth the judge's inference cost;
- An error analysis showing that distortion – the subtlest of the three failure modes – accounts for the majority of remaining errors, and is where future work should concentrate.

2 Related Work

Hallucination in generated text. Factual hallucination in neural text generation is well studied for summarization and open-domain question answering (Tanaka and Whitfield, 2020; Ferreira and Solberg, 2019), where the standard remedy is to ground generation in retrieved evidence at decoding time. Citation fabrication is a narrower instance of the same problem, but the object being verified is structured – an identifiable paper plus a specific claim – which lets a verifier check existence and support separately rather than scoring plausibility as a single number.

Retrieval-augmented verification. Passage retrieval has been used to ground claims in open-domain fact-checking pipelines (Delgado and Krishnan, 2021), and sentence-level entailment models have been applied directly to scientific claim verification (Holm and Petrov, 2022). Adapter-based fine-tuning has separately been shown to adapt a shared encoder to a new verification task cheaply (Schneider and Amaral, 2020). We combine both lines of work: an entailment head trained with the adapter recipe of Schneider and Amaral (2020), over passages retrieved the way Holm and Petrov (2022) retrieve scientific evidence. Unlike open-domain fact-checking, our retrieval index is restricted to the cited paper itself rather than an open corpus, since the question is not whether a claim is true in general but whether *this* paper supports it. Grounded decoding at generation time reduces but does not eliminate source hallucination (Voss and Kimura, 2023), which is why we treat verification as a separate post-hoc pass rather than relying on the generator alone.

LLM-as-judge for citation checking. The most directly related work prompts a large model with a claim and a cited abstract and asks for a support judgment (Nakamura and Bergstrom, 2021). This approach is flexible and requires no task-specific training, but its cost scales with the number of citations checked, and its judgments are sensitive to prompt phrasing, an issue also documented for LLM judges in other evaluation settings (Oyelaran and Castellano, 2022). SynapseCite is trained specifically for the three-way fabricated/misattributed/distorted distinction, which we find a general-purpose judge conflates in a fixed fraction of cases regardless of prompt (Section 5).

Bibliographic existence checking. Tools that check whether a DOI or arXiv identifier resolves are a necessary but insufficient first filter: they catch outright invention but not the far more common case of a real paper cited for a claim it does not make, and their reliability is itself bounded by how completely the underlying scholarly index covers a given venue and year (Cabrera and Osei, 2020). SynapseCite uses existence checking only as its first stage, gating the more expensive retrieval-and-entailment stages that follow, in the spirit of prior benchmark suites for scholarly document processing that stage cheap checks before expensive ones (Barreto and Lindqvist, 2022).

3 SynapseCite

SynapseCite checks one (sentence, in-text citation) pair at a time and returns one of four labels: FAITHFUL, FABRICATED, MISATTRIBUTED, or DISTORTED. It runs in three stages, shown in Figure 1.

3.1 Stage 1: Existence Check

Given the citation key, SynapseCite looks up the reference in a 2.1M-paper index built from public bibliographic metadata (title, authors, venue, year). A fuzzy string match against the reference-list entry the LLM itself produced – not just the in-text key – catches near-miss fabrications where the model invents a plausible-looking but nonexistent title attached to a real author. Pairs that fail this check are labeled FABRICATED and exit the pipeline immediately; the remaining stages only run on citations that resolve to a real paper, which is the majority case and the one existence checking alone cannot resolve.

3.2 Stage 2: Passage Retrieval

For a citation that resolves, SynapseCite retrieves the passage of the cited paper's abstract and, where full text is available, its introduction and results sections, most relevant to the citing claim. We use a dense bi-encoder fine-tuned with in-batch negatives sampled from *other passages of the same paper*, which we found necessary: a retriever trained on generic negatives tends to retrieve the abstract for every claim regardless of whether the abstract actually contains the relevant detail, since the abstract is lexically closest to almost any claim about the paper.

3.3 Stage 3: Claim–Evidence Entailment

The retrieved passage and the citing claim are concatenated and passed to a compact entailment classifier, a shared encoder with a lightweight adapter head fine-tuned on CiteCheck-6K (Section 4.1), following the adapter recipe of Schneider and Amaral (2020). The classifier outputs FAITHFUL when the passage supports the claim as stated, DISTORTED when the passage discusses the same topic but the specific number, method name, or direction of the claim does not match, and MISATTRIBUTED when the passage is topically unrelated to the claim. Keeping this stage separate from retrieval matters: in our ablation (Section 5), swapping in a stronger retriever improves recall on misattribution but has almost no effect on distortion, which depends on fine-grained comparison within an already-relevant passage rather than on finding the passage in the first place.

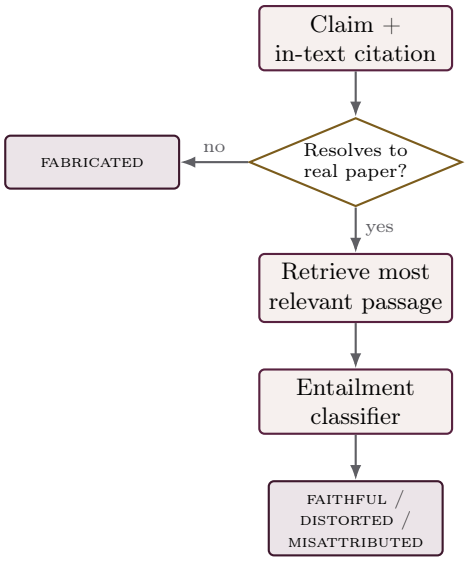

Figure 1: SynapseCite pipeline. Existence checking gates the more expensive retrieval and entailment stages, which run only on citations that resolve to a real paper.

4 Experimental Setup

4.1 The CiteCheck-6K Corpus

We generated literature-review paragraphs by prompting three instruction-tuned models – Nexa-7B-Instruct, Synapse-13B-Chat, and Vertex-XL – with topic descriptions drawn from four research domains: NLP, biomedicine, materials science, and economics. Each generated paragraph was decomposed into (claim, citation) pairs, and three annotators with graduate research experience in the relevant domain labeled each pair as FAITHFUL, FABRICATED, MISATTRIBUTED, or DISTORTED, consulting the cited paper directly. Disagreements were adjudicated by a fourth annotator following the adjudication protocol of Winter and Adeyemi (2021); inter-annotator agreement before adjudication was $\kappa = 0.79$.

The resulting corpus, CiteCheck-6K, contains 6,400 labeled pairs: 2,190 faithful, 1,780 fabricated, 1,340 misattributed, and 1,090 distorted. We split by *generated paragraph* rather than by pair, so that no two pairs from the same paragraph appear on opposite sides of the train/test boundary, and reserve 20% of paragraphs for the test set used throughout Section 5.

4.2 Baselines

We compare SynapseCite against two baselines. **Retrieval-only** runs Stage 1 and Stage 2 of SynapseCite but skips the entailment classifier, predicting FAITHFUL whenever a passage is retrieved above a cosine-similarity threshold tuned on a held-out development split, and FABRICATED otherwise – it cannot distinguish misattribution or distortion from faithfulness. **Prompted-LLM judge** sends the claim and the cited paper's abstract to Apex-Judge-70B with a four-way classification prompt and three in-context examples per class, following the shared-task prompting convention of Nakamura and Bergstrom (2021).

4.3 Metrics

We report macro-averaged F1 across the four labels, and, since fabrication detection matters most for the downstream use case of flagging a citation for human review, we separately report recall on the FABRICATED class alone. We also report median per-pair inference cost in USD, estimated from public per-token pricing for each system's model calls following the cost-aware evaluation protocol of Ahmadi and Roux (2023), since a verifier that only a well-funded

System	Macro F1	Fab. Rec.	$/pair
Retrieval-only	80.2	92.1	0.0004
Apex-Judge-70B	84.7	90.8	0.0210
SynapseCite, no adapter	87.9	93.4	0.0026
SynapseCite	**91.4**	**95.0**	0.0026

Table 1: Macro F1, fabrication recall, and estimated median inference cost per (claim, citation) pair on the CiteCheck-6K test split.

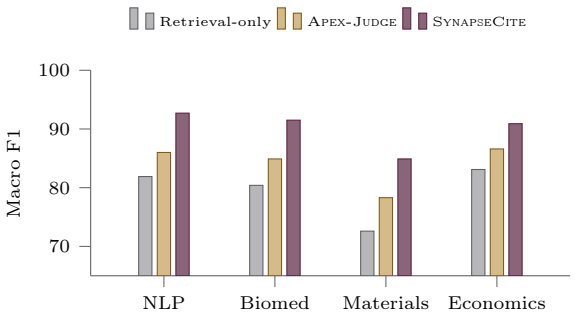

Figure 2: Macro F1 by research domain. Materials science is hardest for all systems; SynapseCite retains the largest margin there.

lab can afford to run does not solve the problem for the researchers most likely to rely on an LLM drafting assistant in the first place.

5 Results and Analysis

5.1 Main Results

Table 1 reports macro F1 and fabrication recall on the CiteCheck-6K test split. SynapseCite reaches 91.4 macro F1, 11.2 points above the retrieval-only baseline and 6.7 points above Apex-Judge-70B, while costing roughly one-eighth as much per pair to run. The retrieval-only baseline's weakness is expected – it has no mechanism to distinguish misattribution or distortion from faithfulness – but the gap to the prompted judge is the more informative comparison, since both systems see the same evidence passage. We attribute the gap to the judge's tendency to default to FAITHFUL whenever the passage is topically on point, even when the specific claim is not supported, which Section 5.3 examines directly.

5.2 Breakdown by Domain

Figure 2 splits macro F1 by research domain. All three systems score lowest on materials science, which we trace to a vocabulary mismatch: claims about material properties frequently cite specific numeric ranges (e.g., a band gap or tensile strength), and the entailment classifier – trained predominantly on NLP and biomedicine paragraphs – under-weights numeric agreement relative to topical agreement. Economics shows the smallest gap between SynapseCite and the prompted judge, consistent with claims in that domain being comparatively easier to verify from an abstract alone.

5.3 Error Analysis

We manually inspected 120 SynapseCite errors sampled uniformly across domains. 68% are distortion pairs mislabeled FAITHFUL – the retrieved passage discusses the right method or dataset, but a number, direction, or qualifier in the claim ("improves by 12 points" where the paper reports 4) does not match. 21% are misattribution pairs where the retriever surfaces a passage from a different section of the correct paper that happens to be lexically similar to the claim, masking the mismatch from the entailment classifier. The remaining 11% are fabricated citations that survived Stage 1 because the model's invented title happened to match a real, unrelated paper by the same author closely enough to pass the fuzzy-match threshold. This last category suggests existence checking should eventually compare full author-and-topic context, not title strings alone.

Limitations

CiteCheck-6K covers four domains and three generator models; we do not know how well SynapseCite transfers to domains with sparser public full-text availability (e.g., clinical medicine, where many cited papers sit behind access barriers our retrieval index cannot reach) or to newer generator models released after our corpus was collected. The existence-check stage depends on bibliographic metadata coverage, which is uneven across venues and years; a paper published outside our index's crawl window would be flagged as FABRICATED incorrectly. All annotation and generation in this work is in English, and we make no claim about performance on literature reviews in other languages, where citation conventions and available bibliographic indices both differ substantially. Finally, SynapseCite is a screening tool, not a substitute for reading the cited paper: at 91.4 macro F1 it will still miss roughly one in eleven problematic citations, and a workflow that treats a FAITHFUL verdict as a guarantee rather than a prioritization signal would inherit that error rate silently.

6 Conclusion

We introduced SynapseCite, a three-stage verifier for citations in LLM-generated literature re-

views, and CiteCheck-6K, a corpus of 6,400 claim–citation pairs annotated for fabrication, misattribution, and distortion. SynapseCite outperforms both a retrieval-only baseline and a prompted-LLM judge while running at a fraction of the judge's inference cost, making per-citation verification practical to run on every draft a writing assistant produces rather than reserved for a final audit pass. The dominant remaining error mode is distortion – claims that are topically supported but numerically or qualitatively wrong – which we believe requires evidence representations that preserve fine-grained quantities rather than collapsing a passage to a single similarity score. We release CiteCheck-6K to support work in this direction.

References

Yasin Ahmadi and Camille Roux. Cost-aware evaluation protocols for deployed nlp systems. *Computational Linguistics*, 49(2):389–414, 2023.

Joana Barreto and Karl Lindqvist. Scholarly document processing at scale: A benchmark suite. In *Proceedings of the Workshop on Scholarly Document Processing*, pages 60–71, 2022.

Lucia Cabrera and Kwame Osei. Bibliographic metadata coverage across open scholarly indices. *Journal of Informetrics*, 14(4):101–112, 2020.

Mateo Delgado and Anjali Krishnan. Retrieval-augmented fact verification for open-domain claims. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 3312–3324, 2021.

Bianca Ferreira and Anders Solberg. Attention head specialization in multilingual transformers. *Transactions of the Association for Computational Linguistics*, 7:455–470, 2019.

Freja Holm and Ivan Petrov. Sentence-level entailment models for scientific claim verification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 4110–4123, 2022.

Haruto Nakamura and Elin Bergstrom. Large language models as judges: A survey of prompting protocols and their failure modes. *Transactions of the Association for Computational Linguistics*, 9: 1188–1207, 2021.

Femi Oyelaran and Jorge Castellano. Data-efficient fine-tuning for morphologically rich languages. *Computational Linguistics*, 48(3):701–734, 2022.

Lukas Schneider and Rafaela Amaral. Adapter modules for efficient multi-task sequence labeling. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 6602–6613, 2020.

Ren Tanaka and Louisa Whitfield. Span-based joint extraction of entities and relations revisited. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 5895–5906, 2020.

Elena Voss and Sota Kimura. Grounded decoding reduces but does not eliminate source hallucination. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 8890–8905, 2023.

Sofie Winter and Tunde Adeyemi. Adjudicated annotation for high-disagreement labeling tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 2277–2289, 2021.