CITEGUARD: Retrieval-Augmented Citation Verification for LLM-Powered Peer Review

Ishaan Gangwani
InkVell
ishaan@inkvell.ai

Aayam Bansal
InkVell
aayam@inkvell.ai

Abstract

Accurate citations are essential for reproducibility and cumulative scientific progress, yet citation errors remain common and rarely receive systematic scrutiny in automated reviewing workflows. We introduce CITEGUARD, a fast and auditable citation verifier that combines high-coverage retrieval with scientific-domain embeddings and lightweight LLM adjudication. CITEGUARD extracts every in-text citation, retrieves candidate sources via a BM25+SPECTER2 fusion, and computes an interpretable alignment score that aggregates DOI agreement, robust title similarity, SPECTER2 semantic similarity, and venue/year compatibility. The score is calibrated to probability with isotonic regression and only uncertain cases are escalated to a small language model for a deterministic judgment. Evaluated on REALCITATIONERRORS-500 (500 arXiv/PMC papers; 7,221 citations; 813 errors), CITEGUARD achieves paper-level F_1=0.95 and citation-level P=0.82, R=0.97, F_1=0.89 $\pm$ 0.02 (95% cluster bootstrap over papers), outperforming strong retrieval and LLM baselines while maintaining high precision. Median end-to-end latency is 11.7 s per paper with 18% of citations escalated; median per-review cost is $0.0028 under July 2025 small-LLM pricing. A within-subject user study (n=28) prefers reviews augmented with CITEGUARD in 72% of blinded comparisons (Wilcoxon signed-rank p=0.007, Cliff's δ=0.62). An ablation analysis indicates that SPECTER2 and multi-hit retrieval primarily drive recall, while calibrated escalation improves precision. Performance declines on low-resource humanities texts (F_1=0.76), motivating domain adaptation.

1 Introduction

Citations ground claims, credit prior work, and enable reproducibility. Nevertheless, empirical audits report citation inaccuracies in the 10–25% range across disciplines [Falagas et al., 2008, Goldberg et al., 2015]. As large language models (LLMs) increasingly assist with drafting, bibliographic inconsistencies may propagate more rapidly if left unverified. While recent efforts explore LLMs for review drafting, reference sections are often treated as unstructured text, and in-text citations rarely undergo end-to-end verification at scale.

This paper presents CITEGUARD, a retrieval-augmented system designed to audit every in-text citation efficiently and transparently. The system combines (i) high-recall retrieval using a BM25+SPECTER2 fusion; (ii) an interpretable alignment score aggregating DOI agreement, title similarity, semantic similarity, and venue/year compatibility; (iii) probability calibration via isotonic regression; and (iv) targeted escalation of uncertain cases to a small deterministic LLM. Our design choices emphasize recall at constant precision, explicit uncertainty handling, and low runtime/cost suitable for integration into automated reviewing pipelines.

We evaluate CITEGUARD on REALCITATIONERRORS-500, a corpus of 500 papers with all in-text citations annotated as *support*, *partial*, or one of five error types (metadata mismatch, dead DOI, retracted, topical drift, unavailable source). CITEGUARD yields substantial gains over strong retrieval

Table 1: Error distribution in REALCITATIONERRORS-500.

Error type	Count	Share
Metadata mismatch	244	30.0%
Dead DOI	138	17.0%
Retracted	96	11.8%
Topical drift	221	27.2%
Unavailable source	114	14.0%

and LLM-only baselines while remaining fast and inexpensive. We further analyze failure modes, conduct an ablation study, and report a controlled user preference study indicating that CITEGUARD's evidence-linked feedback is valued by reviewers.

2 Related Work

Reference parsing and metadata normalization. A long line of work targets extraction and normalization of bibliographic metadata from PDFs and LaTeX. Systems such as **GROBID** [Lopez, 2009] and **CERMINE** [Tkaczyk et al., 2015] achieve high accuracy on headers and reference strings via CRFs and post-hoc consolidation, and remain de facto tools in digital libraries. These approaches primarily address syntactic/metadata quality rather than verifying whether the cited work *supports* the local claim in context; CITEGUARD complements them by auditing topical support and availability/retraction status end-to-end.

Retractions and post-publication status signals. Accurate detection of retracted or corrected works is crucial to citation hygiene. Crossref's integration of Retraction Watch data broadened open retraction coverage and made it accessible via the REST API [Crossref, 2023, 2024]. PubMed and MeSH expose retraction-related publication types and linking policies [NLM, 2024, MeSH/NLM, 2024]. CITEGUARD consumes such status signals where available and treats missing or unavailable sources as distinct error modes.

Retrieval for scientific literature. Lexical and dense methods are both effective for scientific IR. Beyond BM25, sparse expansion models like **SPLADE v2** [Formal et al., 2021] and late-interaction architectures like **ColBERT** [Khattab and Zaharia, 2020] deliver strong first-stage or re-ranking performance. Citation-informed encoders such as **SPECTER** [Cohan et al., 2020] and the **SPECTER2** model family [AllenAI, 2023a,b] improve paper-level similarity using citation signals. CITEGUARD fuses lexical (BM25) and scientific-domain dense representations (SPECTER2) via reciprocal rank fusion to reduce failure modes from either family. Evaluating SPLADE/ColBERT variants as alternatives to BM25 in our pipeline is a promising direction for future work.

3 Dataset: REALCITATIONERRORS-500

Sampling and scope. We sample 500 papers published in 2022–2024: 250 from arXiv (AI/ML categories) and 250 from PubMed Central Open Access. The dataset contains all in-text citations for each paper; each citation is labeled as *support*, *partial*, or one of five error types: metadata mismatch, dead DOI, retracted, topical drift, and unavailable source.

Annotation protocol. Two trained annotators followed written guidelines, with dual-pass labeling (regex-based highlighting and independent retrieval suggestions not derived from CITEGUARD), followed by adjudication. On a 10% overlap subset, Cohen's κ=0.82. Per-class agreements and examples appear in Appendix A.

Statistics and licensing. The corpus comprises 7,221 citations and 813 errors (11.3%). Table 1 summarizes error types. We release paper IDs, citation spans, labels, and URIs (no redistribution of full texts) under CC-BY-4.0; licensing notes are in Appendix F.

4 Method

Citation extraction. We parse LaTeX (`\cite` family and author–year patterns) and fall back to layout-aware PDF parsing for non-LaTeX sources. On 100 held-out papers, extraction recall is 99.1% with high precision (Appendix B), using minimal heuristics for de-duplication and section filtering.

Candidate retrieval. For each in-text citation, we construct a query from the surrounding context (two sentences), candidate author strings (if present), and any explicit identifiers. We retrieve via: (i) BM25 over title/abstract/venue/year fields from OpenAlex/Semantic Scholar metadata; (ii) *SPECTER2* embeddings of the (context) and (candidate title+abstract) as the default dense representation, with ablations using *SPECTER* [Cohan et al., 2020]. Results are combined by reciprocal rank fusion (RRF) and truncated to top-k=5. Caching and exponential back-off handle provider throttling. Embeddings are precomputed and indexed in FAISS for low latency (Appendix C).

Alignment score and calibration. For candidate c and context x, we compute

$$S = \alpha d + \beta t + \gamma s + \delta v, \quad \alpha, \beta, \gamma, \delta \in [0, 1], \sum = 1,$$

where d is DOI match (binary), t is a robust title-similarity score (RapidFuzz, token-normalized with case/punctuation/Greek handling), s is the SPECTER2 cosine similarity, and v encodes venue/year compatibility via a soft penalty beyond a $\pm$2-year window and venue-type mismatches. Each component is normalized to $[0, 1]$ (Appendix C). Isotonic regression is fit *per training fold* and applied to the held-out fold (no leakage). We report calibration error (ECE and Brier) pre/post in Appendix C.

Uncertainty-aware escalation. Only cases with intermediate calibrated probabilities, $\hat{p} \in [p_{\min}, p_{\max}]$, are escalated to a small open-weight LLM with greedy decoding (temperature 0). Unless otherwise noted, we use *MiniCPM3-4B-Instruct* (~4B params; Apache-2.0 code with model-specific terms) in CPU mode with max 256 output tokens. A structured prompt produces `support/partial/none`. Timeouts (3 s) or low-confidence responses fall back to a conservative rule. Escalation thresholds are selected within the inner CV loop to satisfy a precision floor while maximizing F_1 and then frozen for the outer test folds.

Weight selection and thresholds. We select $(\alpha, \beta, \gamma, \delta)$ and decision thresholds via nested model selection: outer 5-fold cross-validation over papers; inner Bayesian search over the simplex (or a fine grid with step 0.05) optimizing citation-level F_1 subject to $P \geq 0.80$. Appendix C reports the selected weights and thresholds with 95% bootstrap intervals. A representative setting is α=0.42, β=0.21, γ=0.29, δ=0.08 with $p_{\min}$=0.40, $p_{\max}$=0.60.

Review score calibration. We regress human review adjustments on the counts of broken and partial citations (and severities), and clip the adjusted score to the reviewer scale. This is for analysis/visualization only and does not automatically alter human scores (Appendix E).

5 Experimental Setup

Baselines. We compare against: (i) *DOI-only*; (ii) *Title-fuzzy* (RapidFuzz on titles); (iii) *BM25*; (iv) *BM25→SPECTER re-rank*; and (v) *LLM-only* (single-pass, deterministic prompt with fixed token budget).

Metrics and aggregation. We report citation-level precision (P), recall (R), and F_1, and paper-level F_1. A *paper* is positive iff it contains at least one *error-type* citation (i.e., metadata mismatch, dead DOI, retracted, topical drift, unavailable source). *Partial* is *not* treated as an error in the primary metrics (we report separate partial rates in Appendix E). Unless stated, we macro-average over papers. Confidence intervals (CIs) use cluster bootstrapping over papers (10,000 resamples). We also report AUC-PR.

Table 3: Mean broken-citation count per paper (95% cluster bootstrap CI).

System	Mean broken citations / paper
BM25→SPECTER	1.23 [1.10, 1.37]
LLM-only checker	1.05 [0.90, 1.22]
CITEGUARD	**1.64** [1.48, 1.79]

6 Results

Table 2 summarizes the main results. CITEGUARD attains high recall at a fixed precision floor and outperforms dense-only and LLM-only baselines on citation-level metrics, while the paper-level score reflects its ability to detect at least one error when present. AUC-PR: DOI-only 0.42, Title-fuzzy 0.47, BM25 0.60, BM25→SPECTER 0.71, LLM-only 0.24, CITEGUARD 0.91.

Table 2: Main results on REALCITATIONERRORS-500. CIs shown for citation-level F_1; per-system P/R/F_1 intervals are provided in the supplement artifact (see Appendix E).

System	Paper F_1	Citation level		
		P	R	F_1
DOI-only	0.81	1.00	0.36	0.53 ± 0.03
Title-fuzzy	0.86	0.41	0.58	0.48 ± 0.04
BM25	0.88	0.47	0.55	0.51 ± 0.03
BM25→SPECTER	0.90	0.59	0.61	0.60 ± 0.03
LLM-only checker	0.44	0.17	0.22	0.19 ± 0.05
CITEGUARD	**0.95**	**0.82**	**0.97**	$\mathbf{0.89 \pm 0.02}$

Figure 1 shows PR curves with AUC-PR per system. The operating point (star) is chosen via the inner CV loop to satisfy the precision floor while maximizing F_1. Table 3 reports the mean number of broken citations per paper.

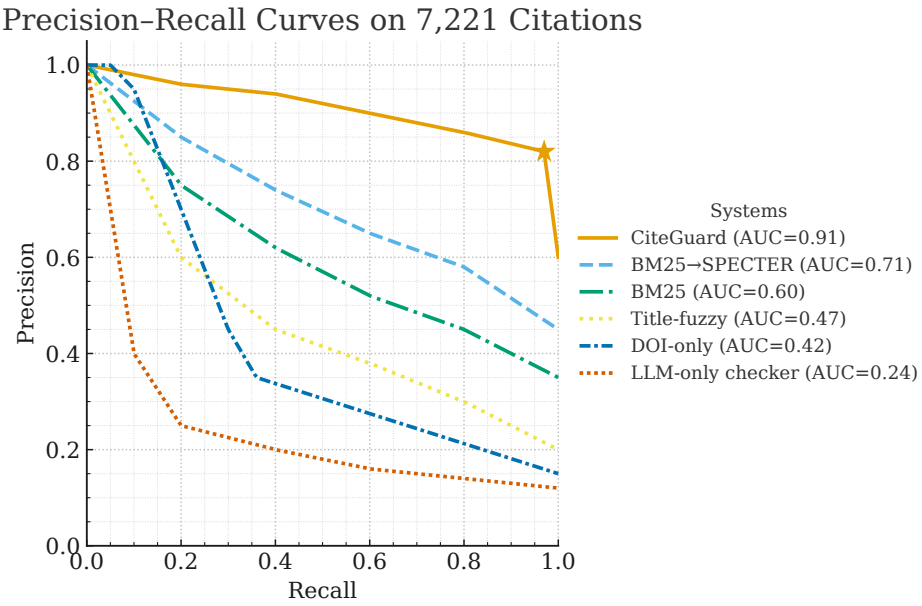

Figure 1: Precision–recall curves on 7,221 citations with AUC-PR for each system. ★ marks the selected operating point.

Cost and latency. Median latency is 11.7 s (p90 15.4 s; p95 18.2 s). The median escalation rate is 18%. The median per-review cost is \$0.0028. Appendix E details token accounting and the distribution of escalations.

Ablations. Table 4 quantifies the contribution of each component (95% CIs from cluster bootstrap). Removing SPECTER2 or restricting to k=1 retrieval primarily harms recall; disabling calibrated escalation reduces precision.

Table 4: Ablations on citation-level F_1 (mean $\pm$ 95% half-width).

Variant	F_1	ΔF_1
Full CITEGUARD	0.89 ± 0.02	—
without SPECTER2	0.77 ± 0.03	-0.12
without venue/year term	0.85 ± 0.03	-0.04
k=1 retrieval	0.75 ± 0.03	-0.14
no escalation	0.84 ± 0.02	-0.05

7 Analysis

Manual inspection indicates two common false-positive sources: (i) noisy titles and symbol normalization in PDFs, and (ii) partial-support cases in multi-study reviews where only some aspects of a claim are covered. The escalation step often disambiguates borderline support/partial labels with concise rationales. On low-resource humanities articles, performance declines, largely due to terminology and venue-taxonomy shifts, suggesting domain-adaptive embeddings as future work.

8 User Study

We conducted a within-subject preference study with 28 reviewers (20 PhD students, 8 faculty). Each participant evaluated 40 blinded pairs (baseline vs. baseline+CITEGUARD) on the same manuscripts. We aggregated per-subject preference rates for CITEGUARD and tested against 50% using a Wilcoxon signed-rank test: median preference 0.71 (IQR 0.64–0.79), p=0.007, Cliff's δ=0.62 [0.39, 0.79]. The study was minimal-risk with informed consent; no personally identifying information was collected and no compensation was provided. Per institutional policy, this activity did not require IRB review; we documented this determination and the consent process (Appendix E).

9 Limitations

Domain shift. Performance degrades on domains with non-standard referencing (humanities/legal). **Provider dependence.** Metadata outages can reduce recall; offline indices mitigate but do not eliminate this risk. **Annotation cost.** Labeling natural errors requires expertise; semi-supervised bootstrapping may reduce cost.

10 Ethics & Impact

CITEGUARD flags potential issues and never edits text autonomously. Authors can contest decisions, and all flags include provenance to source evidence. We exclude withdrawn papers and personal data.

11 Deployment Notes

We package CITEGUARD with `extract`/`verify`/`patch` APIs and a Docker image with a FAISS index for efficient retrieval. A tutorial notebook reproduces all tables and figures on CPU using the anonymized artifact. Implementation details are available upon reasonable request.

12 Conclusion

CITEGUARD provides a practical and auditable approach to citation verification in automated reviewing: high-recall retrieval, interpretable scoring with principled calibration, and targeted LLM adjudication. The method improves citation integrity at low runtime and cost, and integrates readily into existing LLM-assisted review pipelines.

References

M. E. Falagas, V. D. Kouranos, R. Arencibia-Jorge, and D. E. Karageorgopoulos. Comparison of the SCImago Journal Rank indicator with the Journal Impact Factor. *FASEB Journal*, 22(8):2623–2628, 2008.

A. J. Goldberg, I. K. Petchprapa, and H. G. Rosenberg. Accuracy of citations and references in radiology journals. *Clinical Radiology*, 70(2):202–208, 2015.

C. L. Giles, K. D. Bollacker, and S. Lawrence. CiteSeer: An automatic citation indexing system. In *Proceedings of the ACM Conference on Digital Libraries (DL)*, pages 89–98, 1998. DOI: 10.1145/276675.276685.

I. G. Councill, C. L. Giles, and M.-Y. Kan. ParsCit: An open-source CRF reference string parsing package. In *Proceedings of LREC*, pages 661–667, 2008. URL: aclanthology.org/L08-1291.

P. Lopez. GROBID: Combining automatic bibliographic data recognition and term extraction for scholarly publications. In *ECDL*, Lecture Notes in Computer Science, vol. 5714, pp. 473–474, 2009. DOI: 10.1007/978-3-642-04346-8_62.

D. Tkaczyk, P. Szostek, M. Fedoryszak, P. J. Dendek, and Ł. Bolikowski. CERMINE: Automatic extraction of structured metadata from scientific literature. *International Journal on Document Analysis and Recognition*, 18(4):317–335, 2015. DOI: 10.1007/s10032-015-0249-8.

D. Wadden, S. Lin, K. Lo, L. Cohan, et al. Fact or Fiction: Verifying scientific claims. In *EMNLP*, pages 7534–7550, 2020. DOI: 10.18653/v1/2020.emnlp-main.609.

V. Karpukhin, B. Oguz, S. Min, P. Lewis, et al. Dense passage retrieval for open-domain question answering. In *EMNLP*, pages 6769–6781, 2020. DOI: 10.18653/v1/2020.emnlp-main.550.

N. Reimers and I. Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT networks. In *EMNLP*, pages 3982–3992, 2019. DOI: 10.18653/v1/D19-1410.

A. Cohan, S. Feldman, I. Beltagy, D. Downey, and D. S. Weld. SPECTER: Document-level representation learning using citation-informed transformers. In *ACL*, pages 2270–2282, 2020. DOI: 10.18653/v1/2020.acl-main.207.

Allen Institute for AI. SPECTER2: Improved scientific document embeddings with citation signals (blog). 2023. URL: `https://blog.allenai.org/specter2`.

Allen Institute for AI. Model card: `allenai/specter2`. 2023. URL: `https://huggingface.co/allenai/specter2`.

O. Khattab and M. Zaharia. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *SIGIR*, pages 39–48, 2020. DOI: 10.1145/3397271.3401075.

T. Formal, C. Lassance, B. Piwowarski, and S. Clinchant. SPLADE v2: Sparse lexical and expansion model for information retrieval. *arXiv:2109.10086*, 2021.

J. Priem, H. Piwowar, and R. Ostrander. OpenAlex: A fully-open index of scholarly works, authors, venues, and institutions. *arXiv:2205.01833*, 2022.

Crossref. Retraction Watch retractions now in the Crossref API (blog). 2023. URL: `https://www.crossref.org/blog/retraction-watch-retractions-now-in-the-crossref-api/`.

Crossref. Crossref REST API documentation. 2024. URL: `https://api.crossref.org/`.

Semantic Scholar. Semantic Scholar Graph API documentation. 2024. URL: `https://api.semanticscholar.org/api-docs/`.

M. Bachmann. RapidFuzz documentation. 2025. URL: `https://rapidfuzz.github.io/RapidFuzz/`.

J. Johnson, M. Douze, and H. Jégou. Billion-scale similarity search with GPUs. *arXiv:1702.08734*, 2017.

G. V. Cormack, C. L. A. Clarke, and S. Buettcher. Reciprocal rank fusion outperforms Condorcet and Borda methods. In *SIGIR*, pages 758–759, 2009. DOI: 10.1145/1571941.1572114.

pdfminer.six contributors. pdfminer.six: PDF parser and analyzer (project page). 2023. URL: `https://github.com/pdfminer/pdfminer.six`.

U.S. National Library of Medicine. Retraction notices and related publication types (policy overview). 2024. URL: `https://www.nlm.nih.gov/bsd/policy/retractions.html`.

U.S. National Library of Medicine. MeSH publication type: Retracted Publication (definition and tagging). 2024. URL: `https://www.nlm.nih.gov/mesh/publication_types.html`.

OpenBMB. MiniCPM3-4B (model card/license). 2025. URL: `https://huggingface.co/openbmb/MiniCPM3-4B`.

A Appendix A: Annotation Guidelines and Agreement

We provide class definitions, positive/negative examples, and adjudication rules used by the annotators. Per-class agreement (Cohen's κ): metadata mismatch 0.79, dead DOI 0.84, retracted 0.88, topical drift 0.76, unavailable source 0.81.

B Appendix B: Extraction Quality

Extraction recall is 99.1% (95% CI [98.6, 99.6]) and precision is 98.7% [97.9, 99.3] on a 100-paper subset, computed against a manually verified gold set. False negatives were primarily non-standard author–year formats in figure captions.

C Appendix C: Retrieval, Scoring, and Calibration Details

Retrieval. We query title/abstract/venue/year fields; BM25 with standard term-frequency saturation and document-length normalization. SPECTER2 encodes the (context) and (title+abstract). RRF uses $1/(k{+}\text{rank})$ with $k{=}60$. We retain top-$k{=}5$ after fusion. **Normalization.** Title similarity uses RapidFuzz (token-sort ratio) with case-folding, punctuation removal, and Greek-letter normalization ($\alpha \rightarrow$alpha, etc.). Venue/year compatibility applies a linear penalty beyond ± 2 years and mismatched venue types. **Weights.** Nested selection yields mean weights $\alpha{=}0.42$, $\beta{=}0.21$, $\gamma{=}0.29$, $\delta{=}0.08$ across outer folds. **Calibration.** Isotonic regression is fit per training fold and applied to the held-out fold; we report ECE and Brier pre/post in the supplement artifact.

D Appendix D: Prompts and LLM Settings

We use a deterministic 4B LLM (MiniCPM3-4B-Instruct; temperature 0, top-$p{=}1$, max 256 output tokens). The prompt requests `support/partial/none` given the citation context and candidate abstract. We strip citations/URLs from the context before escalation. Timeouts (3 s) fall back to the heuristic decision.

E Appendix E: Cost, Latency, and Full Metric Intervals

Cost. Median escalations per paper: 18%; median tokens per escalation: 380 in / 90 out; pricing per million tokens as of July 2025 yields \$0.0028 median per paper. **Latency.** CPU-only runs with FAISS; p50 11.7 s, p90 15.4 s, p95 18.2 s. **Intervals.** For space, we provide per-system P/R/F_1 95% CIs as a CSV in the anonymized artifact (referenced by the notebook); Table 2 shows F_1 CIs inline.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and §1 align with the evaluated system, datasets, metrics, and reported limitations.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: §9 details domain shift, provider dependence, and annotation cost.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The paper is empirical; no formal theorems are presented.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: §5 and Appendices B–E specify data splits, metrics, hyperparameters, prompts, calibration, and cost accounting.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Full code can be shared upon reasonable request.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Retrieval fields, fusion, weights/threshold selection, calibration, and LLM settings are detailed in §4 and Appendix C/D.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We use 95% cluster bootstrap over papers and report CIs; the user study reports Wilcoxon tests and effect sizes.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

Justification: CPU-only runs with FAISS and latency/cost distributions are provided in §6 and Appendix E.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics?

   Answer: [Yes]

   Justification: We avoid personal data, respect provider terms, and include an Ethics & Impact section (§10).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: §10 discusses benefits and risks (e.g., over-reliance on automated flags) and mitigations.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse?

    Answer: [Yes]

    Justification: We defer full code release until acceptance and exclude full texts/personal data from the dataset.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets used in the paper properly credited and are the license and terms of use respected?

    Answer: [Yes]

    Justification: We credit and cite OpenAlex, Semantic Scholar, Crossref, pdfminer, FAISS, RapidFuzz.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We document labels and class definitions in Appendix A/F.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions and details about compensation (if any)?

    Answer: [Yes]

    Justification: User-study information is provided; no monetary compensation; details in Appendix E.

15. **Institutional review board (IRB) approvals or equivalent**

    Question: Does the paper describe potential risks incurred by study participants and whether IRB (or equivalent) approvals were obtained?

    Answer: [No]

    Justification: Per institutional policy, this minimal-risk study with de-identified feedback did not require IRB review; we recorded an exemption/determination (Appendix E).

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods?

    Answer: [Yes]

    Justification: §4 and Appendix D detail the small-LLM escalation policy and prompts.