CiteExtract: A Transparent Evidence Retrieval System for Scholarly Citation Verification

Diyana Muhammed[1(✉)][0009−0006−1178−4535], Gollam Rabby[2][0000−0002−1212−0101], and Sören Auer[1][0000−0002−0698−2864]

[1] TIB – Leibniz Information Centre for Science and Technology, Hannover, Germany
{diyana.muhammed,soeren.auer}@tib.eu
[2] L3S Research Center, Leibniz University Hannover, Hannover, Germany
gollam.rabby@l3s.eu

Abstract. The reliability of scholarly citations depends on both bibliographic accuracy and semantic faithfulness to the original sources. However, verifying these aspects remains a manual and time consuming process for researchers, reviewers, and editors. As scientific output grows and AI-assisted writing becomes common, the risk of incorrect, fabricated, or misrepresented citations increases significantly. We present CiteExtract, an open-source system designed to support human decision making in citation verification. Rather than making automated judgments, CiteExtract focuses on retrieving and presenting high quality evidence. The system decomposes verification into two complementary stages: a structured metadata layer that validates citations against scholarly databases, and a retrieval based semantic layer that extracts relevant passages from the cited papers. For each citation, the system provides matched database records and specific evidence passages, enabling researchers to inspect the underlying material and make informed final judgments. Verification results are delivered through an annotated PDF and an interactive web interface, enabling researchers to review evidence without disrupting their reading workflow. Across a multi-domain evaluation, the system achieves strong accuracy in metadata validation and competitive accuracy in semantic verification across multiple domains. A within subject user study with 18 researchers across five disciplines yielded a mean SUS score of 77.4 and a UEQ-S usefulness and efficiency score of +2.21. All participants reported that CiteExtract substantially improves verification efficiency, increases transparency, and preserves their full control over citation judgments.

Keywords: Citation verification · Retrieval-augmented generation · Research integrity · Digital libraries.

1 Introduction

Citations are a foundational mechanism of scholarly communication. They connect new claims to prior work, support the traceability of scientific arguments, and enable the cumulative construction of knowledge across disciplines. For

these functions to hold, citations must satisfy two conditions: they must be bibliographically correct, and they must accurately represent the findings of the cited source. Maintaining these properties has become increasingly difficult as the scale and pace of scientific publishing continue to grow. At the same time, large-scale AI writing assistance has lowered the cost of producing fluent but unreliable references [15, 17]. Recent studies show that fabricated, incorrect, and semantically misleading citations are no longer isolated anomalies but an increasingly visible problem in scholarly communication [12, 17, 18, 26, 38, 39].

Citation errors take several forms. Some references point to publications that do not exist [36, 38]; others contain fabricated or inconsistent metadata assembled from multiple real papers [2]; and many are bibliographically correct yet still misrepresent what the cited work actually reported [30]. Existing computational systems address these problems only partially. Tools such as CiteAudit [39] and GhostCite [38] verify whether references exist and whether their metadata is consistent with scholarly databases, while SemanticCite [14] retrieves passages from cited documents to assess whether they support the surrounding claim. To our knowledge, existing systems do not combine both forms of verification within a unified reading workflow. Even when automated systems identify suspicious citations, the verification process itself remains largely manual. Researchers must still leave the manuscript, retrieve the cited source, inspect the relevant passages, and determine whether the citation is appropriate. In many cases, the evidence underlying a system decision is not surfaced directly to the user. As citation errors propagate into indexing services, bibliometric analyses, and scholarly discovery systems, they can gradually undermine the reliability of the scholarly record on which digital libraries depend.

To address this gap, we present CiteExtract, an open-source system for evidence-based citation verification integrated into the scholarly reading workflow. It combines a structured metadata layer with a retrieval-augmented semantic layer, grounding every decision in inspectable database records or retrieved passages rather than a black-box verdict (Section 3). Results are surfaced through annotated PDFs and an interactive interface within a single reading environment.

We evaluate CiteExtract on held-out benchmarks covering both metadata and semantic verification, as well as through a within-subject user study involving 18 researchers across five disciplines. On the metadata benchmark, the deterministic pipeline achieves 96.36% accuracy at a fraction of the cost of LLM-plus-search baselines. On the semantic benchmark, retrieval-augmented verification improves accuracy by up to 17 percentage points over non-retrieval conditions, indicating that evidence quality contributes more strongly to performance than model scale alone. The user study yielded a mean SUS score of 77.4, and all participants reported that citation verification became faster and easier while remaining under their control.

Our contributions are as follows:

– CiteExtract, an open-source system that unifies bibliographic and semantic citation verification within a single evidence-grounded workflow, where every

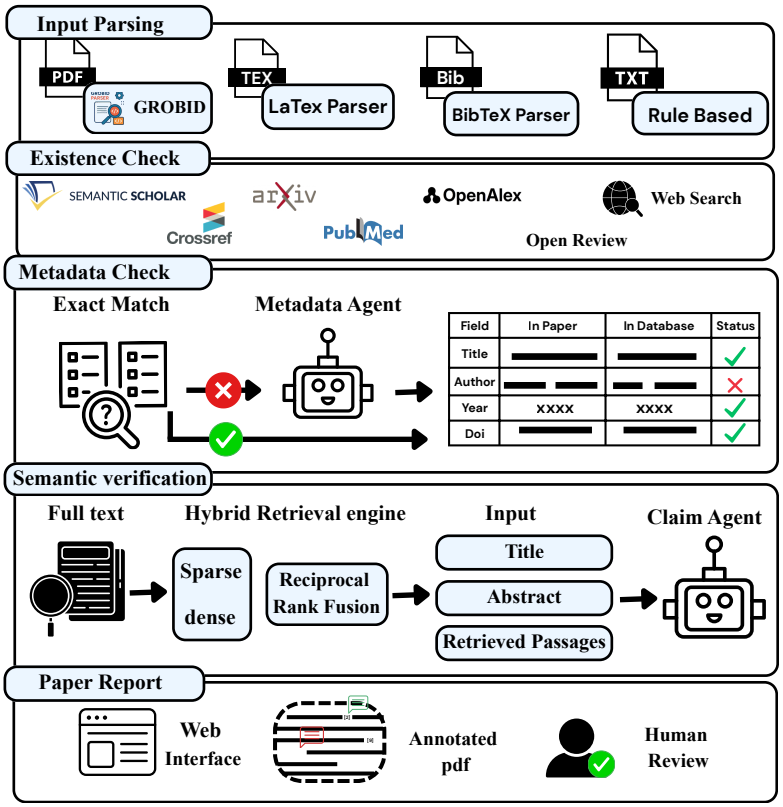

Fig. 1. End-to-end CiteExtract workflow. A submitted document is parsed, each reference is resolved against scholarly databases, metadata fields are validated, and references that pass this stage enter semantic verification. Each citation flag is paired with the evidence passage on which it is based.

decision is traceable to a database record or retrieved passage and can be inspected and overridden directly.

- A reading-integrated interface that supports citation verification and contextual understanding of cited literature within the same environment.
- A comprehensive evaluation spanning metadata verification, semantic verification, and a human-centred usability study, together with released benchmarks and study materials.

The remainder of this paper is organised as follows. Section 2 surveys related work. Section 3 describes the system design and pipeline. Section 4 presents the evaluation, and Section 5 discusses findings, limitations, and directions for future work.

2 Related Work

Citation errors are a long-standing issue in scholarly communication, arising from manual transcription mistakes and the repeated propagation of unverified references [19, 32, 33, 35]. While historically constrained by the pace of human manuscript production, they have become more difficult to detect at scale as scholarly output has increased and citation practices have become increasingly complex. As a result, different types of citation errors now coexist within the same scholarly ecosystem, ranging from incorrect metadata to incorrect or misleading usage of otherwise valid sources. We survey work across three areas: bibliographic and semantic verification, citation attribution, and human-centred scholarly systems, before identifying the gaps that CiteExtract is designed to fill.

2.1 Bibliographic and Semantic Verification

Research on citation verification has largely focused on two complementary problems: validating whether a reference exists and ensuring that it is used correctly in context. Bibliographic verification systems such as CiteAudit [39], GhostCite [38], and CheckIfExist [1] resolve references against scholarly databases to detect fabricated or inconsistent citations. These approaches are effective at identifying non-existent or malformed entries but do not assess whether a cited paper is appropriately used within the text. As a result, citations that are structurally valid but semantically incorrect remain undetected [30].

Semantic verification methods address this limitation directly, asking whether the cited paper actually supports the claim made in the citing sentence. Work in this direction builds on datasets such as SciFact [34], where claims are paired with evidence from scientific documents. SemanticCite [14] retrieves relevant passages from the cited paper and evaluates whether those passages substantiate the claim. FactScore [25] decomposes long-form outputs into atomic claims and verifies each against a retrieval corpus. FACTUM [10] detects citation hallucinations in RAG outputs by analysing a generating model's internal attention pathways, an approach that requires white-box access and cannot be applied to citations in finished manuscripts. Cite Lens [5] detects out-of-scope citations using embedding similarity but does not verify reference existence and does not report precision or recall against a ground-truth benchmark.

Citation attribution and related tasks. A related but distinct line of work studies citation attribution, where the goal is to identify which document a given passage refers to. CiteME [28] introduces a benchmark for this task, while CiteGuard [8] extends it with retrieval-augmented validation. Unlike attribution methods, which infer the referenced source, CiteExtract assumes that the cited document is known and focuses on verifying whether it is used correctly.

2.2 Human-Centred Scholarly Systems

Digital library research has increasingly explored how computational assistance can be embedded directly into scholarly reading workflows, augmenting inline

citations with personalised context, follow-on work, and interactive metadata access [4, 7, 29]. These systems improve discovery, navigation, and contextual understanding of cited literature, but they do not assess whether citations are bibliographically correct or semantically faithful to the cited work. Recent TPDL research has also examined citation behaviour through classification tasks. Duan and Tan [11] distinguish citation intent from cited content type, while Koloveas et al. [21] show that general-purpose LLMs perform competitively on citation classification benchmarks. However, these approaches focus on describing how citations are used rather than verifying whether they are correct.

Taken together, prior work reveals three key limitations. First, bibliographic and semantic verification remain separate concerns, with no unified system addressing both within a single workflow [14, 39]. Second, most systems operate outside the reading environment, requiring users to switch contexts, retrieve sources, and reconstruct evidence manually [7, 27]. Third, verification outputs are typically presented as labels or scores without exposing the underlying evidence needed for human interpretation [10, 14]. These limitations motivate the need for an integrated system that combines verification with evidence presentation directly within the reading workflow. CiteExtract builds on these strands by combining multi-database bibliographic verification, retrieval-based semantic analysis, and human-in-the-loop evidence inspection. Rather than replacing user judgment, it reduces manual effort by surfacing both metadata and supporting passages directly within the manuscript, enabling verification as part of the reading process.

3 The CiteExtract System

Citation verification consists of two complementary tasks: bibliographic validation and semantic evidence checking. The first determines whether a cited work exists and whether its metadata is consistent with scholarly records. The second evaluates whether the cited work actually supports the claim being made. These tasks are fundamentally different in nature: one is a structured retrieval problem over bibliographic databases, while the other requires evidence grounding over full-text content.

CiteExtract explicitly separates these two dimensions and assigns independent labels to each citation: a bibliographic label (Valid, Fabricated, or Unverifiable) and a semantic label (Valid or Misrepresented). A bibliographic citation is Fabricated when no matching scholarly record can be found, Unverifiable when insufficient database coverage prevents a confident determination, and Valid when its metadata is consistent with retrieved records; semantically, a citation is Misrepresented when the retrieved evidence does not support the surrounding claim, and Valid when it does. Every decision is grounded in explicit evidence: database records for bibliographic verification and retrieved textual passages for semantic verification. This design ensures that verification is not a black-box prediction, but an inspectable process in which users can trace and override every decision directly within the document context (Figure 1).

Table 1. Functional (FR) and non-functional (NR) requirements for CiteExtract.

ID	Requirement
Functional Requirements	
FR1	Parse scientific documents (PDF, LaTeX, BibTeX, plain text) and extract in-text citations with surrounding context.
FR2	Resolve references using scholarly databases and classify metadata as Valid, Fabricated, or Unverifiable.
FR3	Retrieve full-text evidence passages and assess whether cited works support claims (Valid vs. Misrepresented).
FR4	Present verification results directly within the reading environment via annotated PDFs and a web interface.
FR5	Support human-in-the-loop correction and export of verification decisions.
Non-Functional Requirements	
NR1	Modular architecture enabling independent replacement of pipeline components.
NR2	Transparency through surfaced evidence, so that every output is traceable to a database record or retrieved passage.
NR3	Reproducibility through released code, prompts, and datasets.
NR4	Low interruption cost: the system must not require users to leave their reading environment to inspect evidence or act on results.

Beyond verification, the same evidence layer also supports contextual reading: by surfacing metadata, abstracts, and retrieved passages alongside each citation, the system reduces the need to manually locate cited works.

3.1 Requirements

Table 1 formalises the functional (FR) and non-functional (NR) requirements derived from Section 2.

3.2 Document Parsing and Citation Linking

CiteExtract supports PDF, LaTeX, BibTeX, and plain-text inputs. PDF documents are processed using GROBID [23], while LaTeX inputs provide direct structured access to citation metadata. All formats are normalised into a unified representation of citation instances, each consisting of a citation marker, its bibliographic entry, and surrounding textual context.

Each citation marker is linked to its corresponding bibliography entry using marker position, pattern matching, and semantic similarity between the citing context and reference titles. This yields a unified set of citation instances that serve as the input for all subsequent verification stages.

3.3 Reference Resolution and Metadata Validation

Each citation instance is resolved against CrossRef, Semantic Scholar, OpenAlex, and PubMed, with additional support for arXiv and OpenReview. Title- and DOI-

Fig. 2. Web interface showing a citation-level validity flag with matched database record, field-level metadata comparison, and retrieval source.

based matching is combined with fuzzy similarity scoring to identify corresponding records. Confidence scores are assigned based on field-level agreement across title, authors, year, and venue, calibrated using isotonic regression to align scores with empirical match rates.

Citations with high confidence database matches are classified directly without LLM involvement, keeping verification cost low for the majority of references. Where database results are ambiguous or conflicting, a lightweight LLM agent reconciles the evidence and produces a structured field-level assessment. Citations are classified as FABRICATED when no reliable match exists, or when metadata is inconsistent with all retrieved candidates. Cases where database coverage is insufficient to reach a confident determination are marked as UNVERIFIABLE. Valid matches proceed to semantic verification.

3.4 Semantic Evidence Verification

For bibliographically valid citations, full text is retrieved by querying open-access repositories, preprint servers, and arXiv in sequence until a source is found; when no full text is available, the citation is marked as UNVERIFIABLE for the semantic stage. Retrieved documents are segmented into overlapping passages and indexed using both sparse (BM25) and dense embeddings. BM25 captures exact lexical matches between claiming sentences and passages, while dense embeddings handle paraphrased or semantically equivalent expressions; the two are fused via Reciprocal Rank Fusion [9] to retain the strengths of both. A cross-encoder reranker then scores the top candidates by joint passage-query relevance, which consistently improves precision over bi-encoder retrieval alone. The final evidence set is evaluated using a structured LLM prompt that determines whether the cited source supports the claim (VALID) or not (MISREPRESENTED), returning both a classification and an explanation grounded in the retrieved passages so that researchers can inspect the reasoning directly.

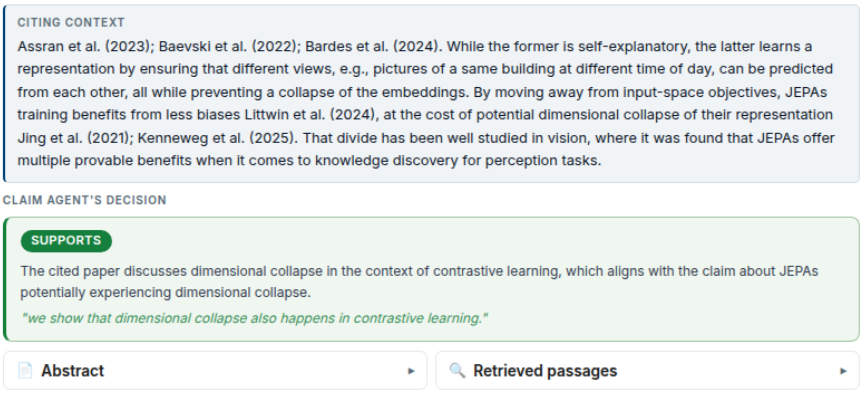

Fig. 3. Web interface for claim verification, showing citing context, cited paper abstract, retrieved passages, and model verdict with reasoning.

3.5 Interfaces and Deployment

CiteExtract provides two tightly integrated interfaces that embed verification directly into scholarly reading (Figures 2 and 3). The web interface presents each citation as an evidence card combining metadata, citation context, retrieved passages, and classification flags, while supporting human override and export of decisions. The annotated PDF preserves the original manuscript while overlaying verification results directly at citation locations. Both interfaces support not only verification but also contextual reading: by exposing abstracts, metadata, and supporting passages at the point of citation, the system enables readers to understand the role of cited work without leaving the document.

CiteExtract is implemented as a modular Python system with a unified API and containerised deployment. Each pipeline component (parsing, retrieval, verification) can be independently replaced, enabling extensibility. The system operates without GPU requirements, making it accessible without specialised infrastructure. Released artefacts include the full pipeline code, prompt templates, hyperparameter configurations, evaluation benchmarks, and user study materials, all available via our GitHub repository.

4 Evaluation

We evaluate CiteExtract along two dimensions: (i) the technical accuracy of its verification pipeline on held-out benchmarks, and (ii) its practical utility when integrated into the reading workflow, including its effect on users' ability to understand and assess cited material directly within context. Beyond binary verification, the system additionally exposes cited titles, abstracts, and retrieved passages, enabling both evidence-grounded checking and rapid contextual understanding of citation relevance.

Table 2. Metadata verification results. V-F1 = F1 on VALID class; F-F1 = F1 on FABRICATED class. Time in minutes; cost in USD for 302 instances; n/a = open-weight model, no API cost. Bold indicates the best-performing CiteExtract configuration.

System	Acc	Macro-F1	V-F1	F-F1	Time	Cost
Llama-3.1-8B	45.03	39.41	57.87	20.95	9.4	n/a
Qwen3-8B	43.38	42.85	48.34	37.36	7.5	n/a
GPT-4o-mini	58.94	58.93	58.39	59.48	8.9	0.03
GPT-4o-mini + search	82.78	82.69	83.95	81.43	43.7	13.41
GPT-4o	61.59	61.53	60.00	63.06	7.9	0.64
GPT-4o + search	84.77	84.76	84.35	85.16	29.2	8.64
GPT-5	68.54	68.50	69.65	67.35	13.6	0.49
GPT-5 + search	88.74	88.66	87.68	89.63	127.4	24.32
GPT-5.5	67.22	65.92	72.58	59.26	92.2	4.65
GPT-5.5 + search	92.38	92.35	91.87	92.83	72.9	33.03
CiteExtract + GPT-4o-mini	96.36	96.36	96.35	96.37	19.9	0.08
CiteExtract + GPT-5	97.02	97.02	97.03	97.01	37.7	0.52
CiteExtract + GPT-5.5	**98.01**	**98.01**	**98.03**	**98.00**	**137.0**	**6.50**

4.1 Datasets

We evaluate CiteExtract on two independent benchmarks corresponding to its two verification stages. The semantic benchmark contains 741 instances, each consisting of a citing sentence, the full text of the cited paper, and a binary label indicating whether the cited work supports the claim (VALID) or misrepresents it (MISREPRESENTED). Instances are drawn from five sources: SemanticCite (266), Sarol et al. (255), CiteME (106), CiteML (71), and CiteBio (43), covering computer science, machine learning, and biomedical literature. All instances have verified full-text access to the cited paper, as retrieval requires the complete document; instances without accessible full text were excluded during benchmark construction. The overall class distribution is 556 VALID (75.0%) and 185 MISREPRESENTED (25.0%); all MISREPRESENTED instances derive from SemanticCite and Sarol et al., while CiteME, CiteML, and CiteBio provide domain coverage across valid citation practices. Given this imbalance, Macro-F1 is the primary metric for semantic evaluation. The metadata benchmark contains 302 balanced instances, each consisting of a reference string and a label indicating whether the reference corresponds to a real publication (VALID) or is fabricated (FABRICATED). Fabricated instances span a range of error types including non-existent references, wrong identifiers, partial authorship errors, and references combining metadata from different real papers. Full construction details and label normalisation mappings are provided in Appendix A.

Evaluation protocol. Semantic evaluation is performed under three evidence conditions: (i) title only, (ii) title plus abstract, and (iii) title, abstract, and retrieved full-text passages. The last condition corresponds to the full CiteExtract pipeline, and also enables downstream interpretability by exposing the evidence

Table 3. Semantic verification accuracy results. V-F1 = F1 on VALID class; M-F1 = F1 on MISREPRESENTED class. Time in minutes.

Model	Condition	Acc	Macro-F1	V-F1	M-F1	Time
Llama-3.1-8B	Title only	56.55	0.559	0.612	0.506	110.8
	Title + Abstract	52.36	0.522	0.547	0.498	60.0
	Title + Abs + Passages	53.98	0.538	0.565	0.512	45.6
Qwen3-8B	Title only	76.11	0.715	0.830	0.601	19.4
	Title + Abstract	71.79	0.683	0.788	0.578	22.7
	Title + Abs + Passages	83.00	0.779	0.885	0.672	26.6
GPT-4o-mini	Title only	72.74	0.698	0.793	0.602	25.9
	Title + Abstract	71.39	0.681	0.784	0.578	25.5
	Title + Abs + Passages	83.81	0.778	0.893	0.663	30.5
GPT-4o	Title only	80.43	0.739	0.870	0.609	16.2
	Title + Abstract	78.00	0.732	0.846	0.618	17.8
	Title + Abs + Passages	85.02	0.789	0.903	0.675	19.2
GPT-5	Title only	81.38	0.770	0.870	0.670	39.3
	Title + Abstract	81.38	0.773	0.869	0.678	40.9
	Title + Abs + Passages	86.37	0.820	0.909	0.731	41.5
GPT-5.5	Title only	69.23	0.669	0.757	0.581	102.1
	Title + Abstract	71.52	0.686	0.782	0.589	79.8
	Title + Abs + Passages	86.10	0.816	0.907	0.725	91.6

used in each decision. Metadata verification compares LLM-only reasoning, LLM with web search, and the CiteExtract pipeline. All experiments use temperature 0; GPT-5 and GPT-5.5 are evaluated at minimal reasoning effort in the main results, with medium reasoning effort results reported in Appendix A for comparison. Full hyperparameter settings including retrieval pool sizes, reranker configuration, and per-model token limits are reported in Appendix D.

4.2 Benchmark Results

Metadata verification. Table 2 reveals a clear pattern: metadata verification is a structured retrieval problem, not a generative reasoning task, and systems designed accordingly perform far better. LLM-only baselines struggle across the board, achieving only 43-62% accuracy, with some models heavily biased toward predicting references as valid and rarely catching fabricated ones. Adding web search helps considerably but at substantial cost and latency. CiteExtract achieves 96.36% accuracy on the balanced benchmark, outperforming the strongest LLM-plus-search baseline by between 4.6 and 13.2 percentage points depending on the comparison. These gains are statistically robust: McNemar tests with Holm-Bonferroni correction confirm significance across all comparisons ($p < 0.01$), and bootstrap analysis verifies that the performance gap does not overlap with any baseline. Notably, the choice of backbone LLM within CiteExtract makes little

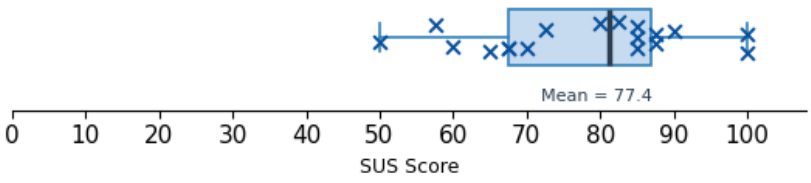

Fig. 4. Distribution of SUS scores across participants (mean 77.4, median 81.2).

difference to the final result: the structured database pipeline drives performance, not the model behind it.

Semantic verification. Table 3 tells a consistent story: what a model has access to matters far more than which model is used. Adding retrieved passages from the cited paper improves accuracy substantially for every model tested, while using only the abstract offers little benefit and can even introduce noise. For GPT-4o-mini, the improvement from retrieval is 11.1 percentage points (72.74% to 83.81%); for GPT-5.5 the gain is even larger at 16.9 points, suggesting that stronger models make better use of good evidence when it is available. Perhaps the most striking finding is that Qwen3-8B, an open-weight model with no API cost, matches GPT-4o-mini when both have access to retrieved passages. This reinforces that retrieval quality is the binding constraint, not model scale. The cost of adding retrieval is also modest: for GPT-4o-mini, it increases cost from $0.17 to $0.25 for the full 741-instance benchmark. Llama-3.1-8B is the exception, showing no improvement and even slight degradation with passages, pointing to weaker instruction-following in evidence-grounded settings. Full per-class F1 scores and medium reasoning effort results are in Table 4.

4.3 User Study

Technical benchmarks establish the reliability of the verification pipeline, but they do not capture how researchers interact with the system during real reading and verification tasks. We therefore conducted a within subject usability study to evaluate how CiteExtract affects the effort, confidence, and workflow involved in citation verification tasks.

Study design and participants. Following the empirical study guidelines of [37] and evaluation conventions from prior TPDL work [20], we recruited 18 participants across five research domains: Biology (9), Physics (3), Computer Science and Information Science (3), Chemistry (2), and Mathematics (1). The study was conducted remotely using a LimeSurvey [22] instrument and followed a three-phase protocol: a baseline phase capturing research background, career stage, and self-reported citation-checking habits; an execution phase in which participants

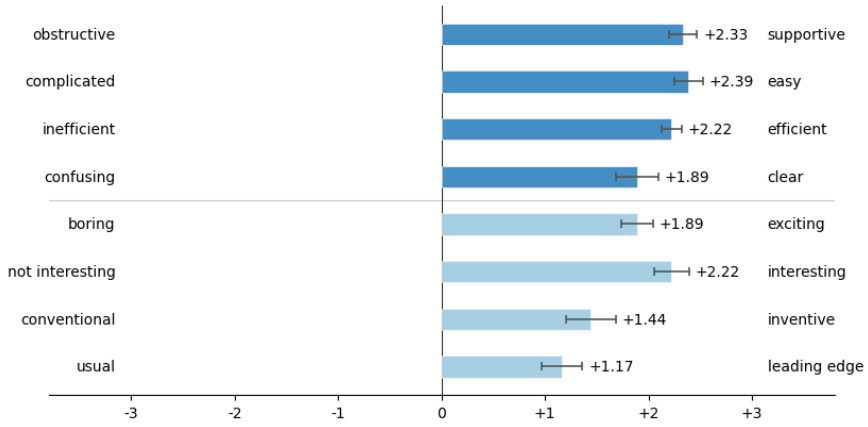

Fig. 5. UEQ-S results across eight adjective pairs on a 7-point scale (−3 to +3). Usefulness and efficiency averaged +2.21; user experience and engagement averaged +1.68.

used CiteExtract on a paper they were writing, reviewing, or reading; and an assessment phase combining standardised usability instruments with open-ended qualitative feedback. Participants reported spending a median of 15 minutes verifying each cited source manually, and 15 of 18 described leaving the reading environment to locate references as highly disruptive, providing a baseline against which system effects were assessed.

Usability and workflow integration. The System Usability Scale (SUS) [3, 6] yielded a mean score of 77.4 (median 81.2, s.d. 14.3), corresponding to a "Good" usability rating (Figure 4). The User Experience Questionnaire Short Form (UEQ-S) [16] showed consistently positive ratings across all eight adjective pairs (Figure 5). The system scored highly in terms of usefulness and efficiency, with an average score of +2.21 out of +3. Participants rated it particularly highly for being "easy" (+2.39) and "efficient" (+2.22). In terms of user experience and engagement, the average score was +1.68, indicating positive user experience alongside strong practical usefulness.

These usability findings were reflected in participants' reported workflow experience (Figure 6). All 18 participants agreed or strongly agreed that CiteExtract saved time compared to their usual citation checking process and made verification easier overall. 17 of 18 participants reported that the citation flags helped them prioritise which references required closer inspection. Importantly, all 18 agreed that they remained in control throughout the process, indicating that the system functioned as assistive support rather than as an automated replacement for expert judgment. 16 of 18 participants agreed that the annotated PDF helped them understand cited papers without leaving their primary document, and 16 of 18 reported that the citation flags helped them understand why each citation was flagged. Open-ended responses highlighted transparency as the most valued

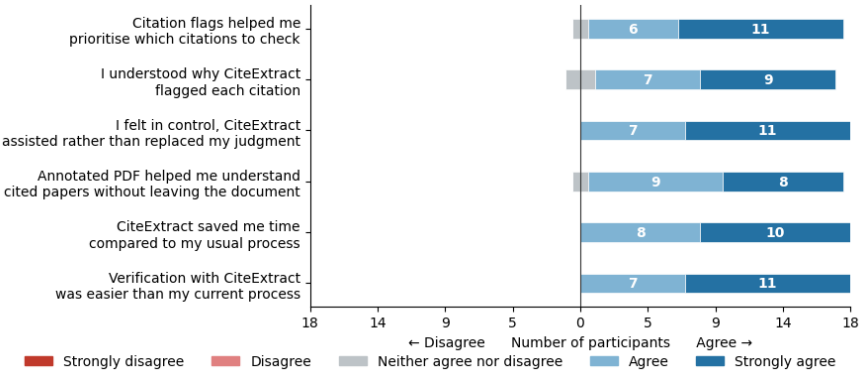

Fig. 6. Participant responses to Likert statements on usability and workflow impact (N=18; 5-point scale: strongly disagree to strongly agree).

aspect, with participants noting that seeing the specific retrieved passage, rather than only a verdict label, was what allowed them to form their own judgment. All 18 participants identified the tool as useful in all three scholarly roles: as authors checking their own work, as reviewers assessing others' manuscripts, and as readers of published literature. Overall, these results indicate that CiteExtract is perceived not only as improving efficiency, but also as strengthening transparency and supporting independent expert judgment across different scholarly workflows.

Real-world validation. As an exploratory real-world validation, we applied CiteExtract to 20 randomly sampled NeurIPS 2025 accepted papers. Each paper had approximately 60 references, making a total of around 1193 references to be reviewed. The system flagged 18 citations for human review, of which 2 were confirmed as genuine semantic misattributions after manual inspection. The remaining 1,175 references were not manually re-verified, so this analysis establishes precision on the flagged subset but not recall. Although limited in scale, this analysis illustrates two complementary points. First, semantic misattribution can persist even in highly selective publication venues. Second, the flag rate of 18 citations across 1193 references reflects the system's design intent: rather than withholding judgement until high confidence is reached, CiteExtract surfaces borderline cases for human inspection, allowing researchers to quickly assess the retrieved evidence and dismiss acceptable citations themselves. Under this design, a higher recall with moderate precision is preferable to conservative flagging that risks missing genuine errors. The confirmed cases are detailed in Appendix C.

5 Discussion

The evaluation highlights that metadata verification and semantic verification require fundamentally different strategies. Metadata validation behaves primarily

as a structured retrieval and consistency-checking task, where deterministic database resolution substantially outperforms purely generative approaches in both accuracy and efficiency. In contrast, semantic verification depends primarily on access to specific supporting evidence. Across all evaluated models, retrieved full-text passages consistently produced large improvements, whereas abstracts alone provided limited benefit and occasionally introduced additional noise. These results suggest that progress in semantic citation verification is likely to depend more on retrieval quality and evidence selection than on scaling reasoning models alone. The findings also indicate that citation verification should be treated as an assistive scholarly workflow rather than a fully automated classification task. Participants consistently relied on the surfaced evidence passages to interpret and assess flagged citations themselves, and several reported that the retrieved context was more useful than the classification label alone. This behaviour reinforces the role of verification systems as tools for evidence inspection and contextual reading, particularly in settings where citation usage may be nuanced or partially subjective. The user study further demonstrates the practical importance of reducing workflow interruption during verification. Participants reported that integrating citation checking directly into the reading environment simplified the process of locating and assessing supporting material. Rather than requiring repeated transitions between databases, PDFs, and search engines, the system consolidates the relevant information around each citation into a single interaction point, reducing the overhead associated with manual verification while preserving researcher control over final decisions.

Taken against the requirements defined in Section 3.1, the evaluation confirms that FR1-FR4 and NR1-NR3 are fully addressed. FR5 (human override and export) and NR4 (low interruption cost) are supported by the interface design and the outcomes of the user study. Multilingual support and retrieval coverage for paywalled documents remain partially addressed and represent important directions for future development.

Ethical Consideration. The study was conducted in accordance with GDPR (Regulation (EU) 2016/679) and the European Code of Conduct for Research Integrity; full data collection and consent procedures are provided in Appendix B. CiteExtract is designed as a decision-support tool, not an automated arbiter of citation correctness - every output is grounded in inspectable evidence and users can override any decision, consistent with the human-oversight principle under the EU AI Act (Regulation (EU) 2024/1689). Commercial LLM APIs are used for a subset of verification steps; prompts contain only bibliographic metadata and short excerpts from openly accessible works, and no personal data is transmitted. Both benchmarks draw exclusively on openly licensed literature; all benchmark data are released under open licences.

Limitations. The most significant structural constraint is retrieval coverage: semantic verification requires full-text access, and for paywalled publications the system falls back to title and abstract, which our results show to be substantially weaker evidence. Affected citations are marked UNVERIFIABLE rather than silently

degraded, but the gap remains. The semantic benchmark covers computer science, machine learning, biomedical, and cancer biology literature; citation conventions in the humanities, law, and social sciences differ substantially, and the binary VALID/MISREPRESENTED scheme may not capture more interpretive citation practices in those fields. The metadata benchmark covers five fabrication types but not subtle misattributions, such as citing a real but topically adjacent paper that does not directly support the specific claim. Metadata confidence scores are calibrated on the benchmark distribution and may drift under systematic shifts in citation style or database coverage in deployment. The user study sample of 18 graduate students, predominantly from biology domain limits generalisability; a larger and more diverse study is needed to confirm that the efficiency and usability benefits extend broadly. Finally, the pipeline assumes English-language input throughout; non-English manuscripts and citations to non-English sources are not reliably handled.

Future work. Future work includes improving retrieval coverage for non-open-access literature, extending support to multilingual and low-resource settings, and incorporating user corrections into iterative refinement of retrieval and verification components. Integrating citation verification into collaborative authoring and peer-review systems represents a promising direction for reducing downstream propagation of citation errors.

6 Conclusion

This paper presented CiteExtract, an open source system for evidence-grounded citation verification that combines structured metadata validation with retrieval-augmented semantic verification within a unified reading workflow. Through annotated PDFs and an interactive interface, the system links citations directly to matched scholarly records and supporting evidence passages, enabling researchers to inspect and assess citations without leaving the document context. Evaluation results confirm that structured database resolution and retrieval quality are the binding constraints in their respective verification stages. A user study further indicates that integrating verification into the reading process improves efficiency and usability while preserving researcher control over final judgments. Together, these findings position citation verification not only as an error-detection task but as a broader problem of evidence access and scholarly reading support.

Acknowledgments. The authors would like to thank the participants of our user study for their time and valuable feedback. This work was conducted as part of the NFDI4DataScience project (DFG project no. 460234259) and the KISSKI AI Service Center (01IS22093C), funded by the German Ministry of Research, Technology and Space (BMFTR).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

Bibliography

[1] Abbonato, D.: Checkifexist: Detecting citation hallucinations in the era of ai-generated content. arXiv preprint arXiv:2602.15871 (2026)

[2] Ansari, S.: Compound deception in elite peer review: A failure mode taxonomy of 100 fabricated citations at neurips 2025. arXiv preprint arXiv:2602.05930 (2026)

[3] Bangor, A., Kortum, P.T., Miller, J.T.: An empirical evaluation of the system usability scale. International Journal of Human–Computer Interaction **24**(6), 574–594 (2008)

[4] Basti, Y.A., Davoudi, H., Neshati, A.: Citepeek: Contextual citation exploration within research papers. In: Proceedings of the 7th ACM Conference on Conversational User Interfaces. pp. 1–7 (2025)

[5] Broise, J.B.d.l., Sauerburger, F., Sayas, E., Tecu, D.M., Meijere, S., Cuculovic, M.: Cite lens: An ai tool for detecting out-of-scope and out-of-context citations. In: International Conference on Theory and Practice of Digital Libraries. pp. 25–34. Springer (2025)

[6] Brooke, J.: SUS: A quick and dirty usability scale. Usability Evaluation in Industry **189**(194), 4–7 (1996)

[7] Chang, J.C., Zhang, A.X., Bragg, J., Head, A., Lo, K., Downey, D., Weld, D.S.: Citesee: Augmenting citations in scientific papers with persistent and personalized historical context. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. pp. 1–15 (2023)

[8] Choi, Y.M., Guo, X., Fung, Y.R., Wang, Q.: Citeguard: Faithful citation attribution for llms via retrieval-augmented validation. arXiv preprint arXiv:2510.17853 (2025)

[9] Cormack, G.V., Clarke, C.L.A., Buettcher, S.: Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 758–759. ACM (2009)

[10] Dassen, M., Kotula, R., Murray, K., Yates, A., Lawrie, D., Kayi, E., Mayfield, J., Duh, K.: Factum: Mechanistic detection of citation hallucination in long-form rag. arXiv preprint arXiv:2601.05866 (2026)

[11] Duan, C., Tan, Z.: Semantically orthogonal framework for citation classification: Disentangling intent and content. In: International Conference on Theory and Practice of Digital Libraries. pp. 183–206. Springer (2025)

[12] Emsley, R.: Chatgpt: these are not hallucinations–they're fabrications and falsifications. Schizophrenia **9**(1), 52 (2023)

[13] Esau, P., Shmatko, N., Adam, A.: GPTZero finds over 50 new hallucinations in ICLR 2026 submissions. https://gptzero.me/news/iclr-2026/ (2026), accessed: 2026-05-13

[14] Haan, S.: Semanticcite: Citation verification with ai-powered full-text analysis and evidence-based reasoning. arXiv preprint arXiv:2511.16198 (2025)

[15] Hanson, M.A., Barreiro, P.G., Crosetto, P., Brockington, D.: The strain on scientific publishing. Quantitative Science Studies **5**(4), 823–843 (2024)
[16] Hinderks, A., Schrepp, M., Thomaschewski, J.: Design and evaluation of a short version of the user experience questionnaire (UEQ-S). International Journal of Interactive Multimedia and Artificial Intelligence **4**(6), 103–108 (2017)
[17] İlter, H.K.: The 17% gap: Quantifying epistemic decay in AI-assisted survey papers. arXiv preprint arXiv:2601.17431 (2026)
[18] Jain, A., Nimonkar, P., Jadhav, P.: Citation integrity in the age of AI: evaluating the risks of reference hallucination in maxillofacial literature. Journal of Cranio-Maxillofacial Surgery (2025)
[19] Jergas, H., Baethge, C.: Quotation errors in general medicine journals. PeerJ **3**, e1364 (2015)
[20] John, L., Ghanmi, A.M., Wittenborg, T., Auer, S., Karras, O.: Extractable: Human-in-the-loop transformation of scientific corpora into structured knowledge. In: International Conference on Theory and Practice of Digital Libraries. pp. 470–487. Springer (2026)
[21] Koloveas, P., Chatzopoulos, S., Vergoulis, T., Tryfonopoulos, C.: Can llms predict citation intent? an experimental analysis of in-context learning and fine-tuning on open llms. In: International Conference on Theory and Practice of Digital Libraries. pp. 207–224. Springer (2025)
[22] LimeSurvey GmbH: LimeSurvey: Open source survey tool (2024), https://www.limesurvey.org, version 6.x
[23] Lopez, P., et al.: GROBID: A machine learning software for extracting information from scholarly documents (2008), https://github.com/kermitt2/grobid, version 0.8.0
[24] Magar, I., Schwartz, R.: Data contamination: From memorization to exploitation. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 157–165 (2022)
[25] Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W.t., Koh, P., Iyyer, M., Zettlemoyer, L., Hajishirzi, H.: Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 12076–12100 (2023)
[26] Naser, M.: How llms cite and why it matters: A cross-model audit of reference fabrication in ai-assisted academic writing and methods to detect phantom citations. arXiv preprint arXiv:2603.03299 (2026)
[27] Oelen, A., Jaradeh, M.Y., Auer, S.: Orkg ask: a neuro-symbolic scholarly search and exploration system. arXiv preprint arXiv:2412.04977 (2024)
[28] Press, O., Hochlehnert, A., Prabhu, A., Udandarao, V., Press, O., Bethge, M.: Citeme: Can language models accurately cite scientific claims? Advances in Neural Information Processing Systems **37**, 7847–7877 (2024)
[29] Rachatasumrit, N., Bragg, J., Zhang, A.X., Weld, D.S.: Citeread: Integrating localized citation contexts into scientific paper reading. In: Proceedings of the 27th International Conference on Intelligent User Interfaces. pp. 707–719 (2022)

[30] Sarol, M.J., Ming, S., Radhakrishna, S., Schneider, J., Kilicoglu, H.: Assessing citation integrity in biomedical publications: corpus annotation and nlp models. Bioinformatics **40**(7), btae420 (2024)
[31] Shmatko, N.: GPTZero finds 100 new hallucinations in NeurIPS 2025 accepted papers. https://gptzero.me/news/neurips/ (2026), accessed: 2026-05-13
[32] Siebers, R., Holt, S.: Accuracy of references in five leading medical journals. The Lancet **356**(9239), 1445 (2000)
[33] Simkin, M.V., Roychowdhury, V.P.: Read before you cite! arXiv preprint cond-mat/0212043 (2002)
[34] Wadden, D., Lin, S., Lo, K., Wang, L.L., van Zuylen, M., Cohan, A., Hajishirzi, H.: Fact or fiction: Verifying scientific claims. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 7534–7550 (2020)
[35] Wager, E., Middleton, P.: Technical editing of research reports in biomedical journals. Cochrane Database of Systematic Reviews (2) (2007)
[36] Walters, W.H., Wilder, E.I.: Fabrication and errors in the bibliographic citations generated by chatgpt. Scientific Reports **13**(1), 14045 (2023)
[37] Wohlin, C., Runeson, P., Höst, M., Ohlsson, M.C., Regnell, B., Wesslén, A.: Empirical strategies. In: Experimentation in Software Engineering, pp. 9–36. Springer (2012)
[38] Xu, Z., Qiu, Y., Sun, L., Miao, F., Wu, F., Wang, X., Li, X., Lu, H., Zhang, Z., Hu, Y., et al.: Ghostcite: A large-scale analysis of citation validity in the age of large language models. arXiv preprint arXiv:2602.06718 (2026)
[39] Yuan, Z., Shi, K., Zhang, Z., Sun, L., Chawla, N.V., Ye, Y.: Citeaudit: You cited it, but did you read it? a benchmark for verifying scientific references in the llm era. arXiv preprint arXiv:2602.23452 (2026)

A Benchmark Construction Details

Semantic Benchmark

Each source uses a different labelling scheme, which was mapped to the binary VALID / MISREPRESENTED scheme as follows.

SemanticCite [14] uses a four-class scheme: support, partial-support, no-support, and contradicts. Support was mapped to VALID; no-support and contradicts were mapped to MISREPRESENTED. Partial-support instances did not map cleanly to either class and were excluded.

Sarol et al. [30] use an accuracy-classification scheme assessing whether a citation faithfully represents the cited work. Accurate citations were mapped to VALID; misrepresenting or distorting citations were mapped to MISREPRESENTED. Instances labelled as unverifiable were excluded.

CiteME [28] is a citation attribution dataset in which the task is to identify which paper a passage cites. To repurpose it for semantic verification, instances where the gold-cited paper substantiates the content of the citing passage were

labelled VALID. Instances where no clear support relationship could be established were excluded.

We construct two domain-specific datasets, CiteML (machine learning) and CiteBio (cancer biology), following the curation methodology of Press et al. [28]. For each domain, papers were selected and citation excerpts manually extracted that make specific, verifiable claims and provide sufficient context to identify the cited work. Each instance records the source paper, the citation context, and the ground-truth cited paper. To ensure evaluation measures retrieval capability rather than memorisation, we apply memory filtering using GPT-4o in a zero-shot setting with no external tools: instances where GPT-4o correctly identifies the cited paper from context alone are removed [24]. Following full-text availability filtering, required because semantic verification depends on access to the complete cited document, CiteML retains 71 instances and CiteBio retains 43 instances. Both datasets contain only VALID instances by construction, as excerpts were selected to contain specific, attributable claims with clear supporting evidence in the cited work. CiteML and CiteBio are released as part of the CiteExtract repository.

Metadata Benchmark

The metadata benchmark contains 151 fabricated references drawn from NeurIPS 2025 accepted papers (100) and ICLR 2026 submissions (51), and 151 valid references sampled from Semantic Scholar across multiple domains. Fabricated instances were drawn from independently verified sources: 100 hallucinated citations originally identified by GPTZero [31] across 53 NeurIPS 2025 accepted papers and subsequently taxonomised by Ansari [2], and 51 hallucinated citations identified in ICLR 2026 submissions by GPTZero [13]. Both sets were verified against real publication records prior to inclusion; no instance was labelled by the system under evaluation.

Analysis of the verification notes reveals a taxonomy of fabrication types present in the benchmark:

- **Pure non-existence**: the reference corresponds to no real publication.
- **Wrong identifier**: the title and authors are plausible but the DOI, arXiv ID, or URL is fabricated or misassigned.
- **Fabricated references**: metadata from different real papers is combined into a single reference (e.g., authors from one paper, title from another).
- **Partial authorship fabrication**: the paper exists but some or all authors are incorrect.
- **Wrong venue or year**: the paper exists but is attributed to the wrong journal, conference, or publication year.

Tables 4 and 5 report results at medium reasoning effort for GPT-5 and GPT-5.5, complementing the minimal effort results in the main paper.

Table 4. Semantic verification results for GPT-5 and GPT-5.5 at medium reasoning effort across three evidence conditions.

Model	Condition	Acc	Macro-F1	V-F1	M-F1	Time
GPT-5 (med)	Title only	69.64	0.666	0.767	0.565	176.0
	Title + Abstract	72.60	0.693	0.794	0.593	139.8
	Title + Abs + Passages	82.86	0.776	0.885	0.667	157.5
GPT-5.5 (med)	Title only	70.99	0.681	0.777	0.586	124.7
	Title + Abstract	73.01	0.699	0.796	0.603	131.3
	Title + Abs + Passages	85.96	0.812	0.907	0.717	110.6

Table 5. Metadata verification results for GPT-5 and GPT-5.5 at medium reasoning effort, including CiteExtract for comparison.

System	Acc	Macro-F1	V-F1	F-F1	Time	Cost
GPT-5 (med)	70.86	70.86	71.05	70.67	108.2	5.20
GPT-5 + search (med)	90.73	90.66	89.86	91.46	313.1	36.92
GPT-5.5 (med)	68.21	66.30	74.33	58.26	194.9	7.09
GPT-5.5 + search (med)	93.38	93.35	92.96	93.75	388.2	45.82
CiteExtract + GPT-5.5 (med)	98.01	98.01	98.03	98.00	137.0	6.50

B User Study Protocol

Participant Overview

Figure 7 summarises the demographic composition of the 18 study participants across research stage and domain.

Ethics and Data Handling

The study was conducted in accordance with institutional data protection guidelines. All responses were collected anonymously via a self-hosted LimeSurvey instance on EU servers, with IP logging and referrer tracking disabled. No personally identifiable information was recorded. Participants provided informed consent prior to participation and were given the option to consent to inclusion of their anonymised responses in a public dataset. The complete survey instrument and anonymised response data are available in the project repository at GitHub.

Survey Structure

The survey comprised three phases administered in sequence. Table 6 summarises the structure, purpose, and item count of each phase.

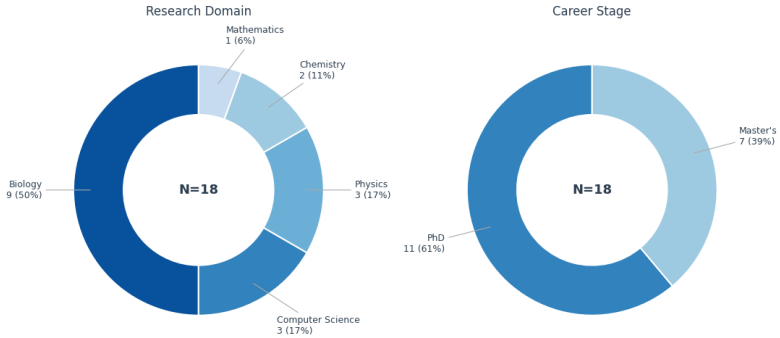

Fig. 7. Participant distribution by career stage (left) and research domain (right).

Table 6. Survey structure by phase.

Phase	Purpose	Instrument	Items
Background	Research field, career stage, LLM usage frequency, citation-checking habits, prior encounters with fabricated or misrepresenting citations, time cost and disruption of manual verification	Custom	9
Experience	UEQ-S usability ratings; custom Likert items on transparency, workflow integration, and control; system adoption questions; open-ended feedback	UEQ-S [16] + Custom	19
Usability	Standard SUS questionnaire	SUS [6]	10

Task Instructions

Participants were asked to use CiteExtract on a scientific paper of their own choosing. Eligible papers included manuscripts they were currently writing, a paper under review, or any published paper they had read recently. No paper was assigned by the study; the choice was left entirely to the participant to ensure ecological validity. Participants were asked to run CiteExtract on the selected paper, inspect at least three flagged citations using both the web interface and the annotated PDF, and form their own judgment on each flag before completing the assessment phase.

Open-Ended Prompts

Table 7 shows representative participant responses to the open-ended prompts on usefulness and frustration.

Table 7. Selected participant responses to open-ended survey prompts.

What was the most useful thing CiteExtract did for you?	
1	*"Minimised the time spent searching for the accuracy of citations, enhancing efficiency in the reading process."*
2	*"The transparency, rather than telling it is correct or not."*
3	*"I didn't have to search different sources and verify my references, I could easily check all in one place."*
4	*"Easy to gather more information on cited paper."*
What was most frustrating or confusing?	
1	*"It didn't actually provide claim verdicts for all citation contexts."*
2	*"To upload PDF files."*

C Real-World Deployment

CiteExtract flagged 18 citations as potentially misrepresenting their sources; following manual human verification, 2 were confirmed as genuine semantic misattributions, while the remaining 16 were judged as acceptable or borderline uses of the cited work. Table 8 summarises the two confirmed cases.

Table 8. Confirmed citation misrepresentations identified by CiteExtract in NeurIPS 2025 accepted papers, validated by human review.

Case 1	
Cited paper	*Orthogonal Gradient Descent for Continual Learning* (Farajtabar et al., 2020)
Citing context	*"Notably, the saliency of weights is determined by analyzing the activation space (sensitivities) of the base model rather than just regularizing based on the weight space, similar to [10, 26]."*
Analysis	OGD operates entirely in gradient/weight space; it performs no activation-space analysis.
Case 2	
Cited paper	*Neural Markov Random Field for Stereo Matching* (Guan et al., 2024)
Citing context	*"The former methods predict the disparity update and iteratively approximate the GT value in a GRU framework [27, 24, 16, 48, 5, 10]. They have achieved great performance in both benchmark and zero-shot generalization testing."*
Analysis	NMRF is an MRF-based method explicitly contrasted with GRU iterative refinement in its own related work.

D Hyperparameters

Table 9. Hyperparameter configuration for the semantic verification stage, covering document chunking, hybrid retrieval, LLM-based verdict generation, and pipeline execution settings used during our experiment.

Component	**Value**
Chunking	
Chunk size	512 characters
Overlap	50 characters
Minimum chunk length	10 words
Retrieval	
BM25 candidates	10
Dense candidates	10
Dense model	`all-MiniLM-L6-v2` (384-dim)
Fusion	Reciprocal Rank Fusion, k=60
Reranker	FlashRank (ms-marco-MiniLM-L-12-v2)
Top passages per query	3
Verification	
Temperature	0
Max output tokens	512
Abstract truncation	1500 characters
Structured output	JSON schema enforced
Runner	
Concurrency	12 simultaneous records
Timeout per request	60 seconds