

# BIBAGENT: An Agentic Framework for Traceable Miscitation Detection in Scientific Literature

Peiran Li<sup>1,2</sup> Fangzhou Lin<sup>1,3</sup> Shuo Xing<sup>1</sup> Xiang Zheng<sup>2</sup> Xi Hong<sup>2</sup> Jiashuo Sun<sup>4</sup>  
 Zhengzhong Tu<sup>1</sup> Chaoqun Ni<sup>2</sup>

## Abstract

Citations are the bedrock of scientific authority, yet their integrity is compromised by widespread miscitations: ranging from nuanced distortions to fabricated references. Systematic citation verification is currently unfeasible; manual review cannot scale to modern publishing volumes, while existing automated tools are restricted by abstract-only analysis or small-scale, domain-specific datasets in part due to the “paywall barrier” of full-text access. We introduce BIBAGENT, a scalable, end-to-end agentic framework for automated citation verification. BIBAGENT integrates retrieval, reasoning, and adaptive evidence aggregation, applying distinct strategies for accessible and paywalled sources. For paywalled references, it leverages a novel *Evidence Committee* mechanism that infers citation validity via downstream citation consensus. To support systematic evaluation, we contribute a 5-category *Miscitation Taxonomy* and MISCITEBENCH, a massive cross-disciplinary benchmark comprising 6,350 miscitation samples spanning 254 fields. Our results demonstrate that BIBAGENT outperforms state-of-the-art Large Language Model (LLM) baselines in citation verification accuracy and interpretability, providing scalable, transparent detection of citation misalignments across the scientific literature.

## 1. Introduction

Citations are central to scientific communication, shaping how knowledge claims are established, credit is assigned, and research is evaluated. As such, citations are widely treated as reliable indicators of conceptual connection and research impact, underpinning formal research evaluations

<sup>1</sup>Texas A&M University <sup>2</sup>University of Wisconsin-Madison  
<sup>3</sup>Worcester Polytechnic Institute <sup>4</sup>University of Illinois Urbana-Champaign. Correspondence to: Peiran Li <[lipeiran@tamu.edu](mailto:lipeiran@tamu.edu)>, Chaoqun Ni <[chaoqun.ni@wisc.edu](mailto:chaoqun.ni@wisc.edu)>.

Preprint. January 27, 2026.

for hiring, promotion, and funding (Waltman, 2016). This system, however, rests on the assumption that citations are accurate. Yet, a growing body of scholarship reveals that citation practices are often inconsistent, flawed, or strategically manipulated (Simkin & Roychowdhury, 2003; Wilhite & Fong, 2012; Greenberg, 2009). This phenomenon, *miscitation*, occurs when cited sources are distorted, fabricated, or taken out of context. The scale is alarming: studies show high rates of inaccuracy, with one analysis finding **39%** of sampled biomedical citations were inaccurate (Sarol et al., 2024), and other manual audits reporting misquotation rates between **10%-25%**, depending on the field (De Lacey et al., 1985; Jergas & Baethge, 2015; Rekdal, 2014). This is not just accidental; the high-stakes evaluation system has spurred deliberate manipulation, including coercive citation (Wilhite & Fong, 2012), citation cartels (Secchi, 2022), and black market schemes (Chawla, 2024). This long-standing problem is now exacerbated by the emergence of generative AI. Large Language Models (LLMs) introduce new risks by producing “*hallucinated citations*”, fabricated or distorted references, and their integration into scholarly workflows threatens widespread, systematic errors (Walters & Wilder, 2023; Liang et al., 2024; Rao et al., 2025). Miscitations have severe consequences: they distort the evidentiary basis of claims, propagate inaccuracies, and erode the reliability of citation-based metrics, thereby undermining the integrity of the scientific record and the fairness of the research ecosystem.

Yet, detecting miscitations is inherently challenging. Even for expert reviewers and editors, verifying citation accuracy requires close reading and detailed familiarity with both the citing and cited texts, including the ability to assess whether the citation faithfully represents the original work in content, context, and intent. Given that the average scientific article cites more than 45 references (Dai et al., 2021), and that such assessments demand considerable expertise and time, *manual verification is not scalable*. The challenge is further compounded by the fact that a substantial portion of scientific literature is behind paywalls or otherwise inaccessible, making it impossible for reviewers to fully evaluate the cited sources.

Automated approaches based on natural language process-

ing (NLP) offer promise, but existing systems face three fundamental limitations. First, the definition of “mispresentation” remains conceptually ambiguous. Traditional NLP approaches have historically oscillated between coarse-grained sentiment analysis (Teufel et al., 2006) and binary fact-checking (Athar & Teufel, 2012). These frameworks ignore nuanced errors, such as scope extrapolation or evidence characterization, in which a citation may be technically “supportive” but methodologically misrepresented. Second, current systems suffer from “data myopia” and an “inaccessibility barrier.” High-performing benchmarks like (Wadden et al., 2020) rely primarily on abstracts, yet the evidence required to identify contradictions often resides in other parts of scientific articles, such as experimental tables or appendices. Furthermore, since much of the scientific record remains behind paywalls, AI systems frequently default to silence or “hallucinate” source content when they cannot access the full text. Finally, the emergence of LLMs has introduced the peril of “sycophancy,” wherein models tend to mirror the user’s claims rather than rigorously challenging them against the evidence (Wei et al., 2024). Collectively, these barriers, compounded by a scarcity of labeled data and disciplinary heterogeneity, have limited research to small, proprietary datasets, leaving the scientific community without a robust, generalizable system for maintaining citation integrity in an era of AI-augmented writing.

To bridge these gaps, we present BIBAGENT, a comprehensive framework designed to evaluate the alignment between citing claims and cited papers for research integrity. Unlike some prior approaches that treat misrepresentation detection as a single classification task, BIBAGENT recognizes that misrepresentation detection is a multi-layered investigative process that must adapt to varying levels of data accessibility. For *accessible sources* (i.e., the full text of a cited paper is accessible), we introduce the *Accessible Cited Source Verifier* (ACSV), which uses an adaptive multi-stage architecture. It moves from efficient bi-encoder retrieval to deep, Large Language Model (LLM)-driven verification only when ambiguity arises. This design ensures traceability and reduces token consumption by **79.4%** without compromising the **100% detection rate**.

The most significant departure from existing methodologies lies in our treatment of the “Inaccessibility Barrier” through the *Inaccessible Cited Source Verifier* (ICSV), in case the full text of a cited paper is inaccessible. Rather than defaulting to silence when faced with paywalled text, ICSV aggregates a “voting committee” of open-access papers that cite the same source. By computing a field-normalized consensus among these downstream citers, BIBAGENT reconstructs the content of hidden sources through collective community intelligence. This shift-from direct document inspection to a multi-perspective consensus model-allows for a robust chain of integrity even when the primary record

is unavailable.

We further introduce two foundational contributions for rigorous evaluation within and beyond the project. First, we formalize a *Unified Taxonomy of Mispresentation*, categorizing errors into five distinct classes ranging from Content Misrepresentation to Scope Extrapolation. Second, we release MISCITEBENCH, a large-scale, cross-disciplinary benchmark of 6,350 expert-validated samples spanning 254 fields, constructed via a “knowledge-blank” protocol to prevent LLM contamination. Together, these resources provide the first robust, generalizable testbed for automated citation verification.

In summary, our core contributions are:

- **BIBAGENT**, the *first* end-to-end agentic framework capable of handling misrepresentations involving both accessible and inaccessible cited sources, offering a robust solution to the paywall problem via a community consensus mechanism.
- **Adaptive verification**: a multi-stage “zoom-in” logic balancing efficiency and precision, with explicit reasoning trails.
- **Evaluation infrastructure**: a comprehensive 5-category *Mispresentation Taxonomy* and MISCITEBENCH, the largest cross-disciplinary benchmark to date, enabling scalable and reproducible assessment of citation integrity.

## 2. Taxonomy and MISCITEBENCH Construction

Automated misrepresentation detection requires a precise characterization of how citations can fail. Prior work largely limited to anecdotal case studies or coarse binary labels (valid vs. invalid) (Pride & Knoth, 2017; 2020), and domain-specific typologies (e.g., in clinical medicine or psychology) built around loosely defined sentiment categories (Wadden et al., 2020; 2022; Wadden & Lo, 2021; Sarol et al., 2024). These schemes are challenging to generalize across fields and fail to provide an operational error code space for large, heterogeneous corpora. To our knowledge, there is still *no* unified taxonomy that (i) is mutually exclusive and collectively exhaustive at the error-code level for multi-field scientific corpora and (ii) can be applied reproducibly across disciplines.

We address this gap by introducing (i) a five-category taxonomy of misrepresentation *errors* governed by simple operational “litmus tests,” and (ii) MISCITEBENCH, a contamination-controlled benchmark that instantiates this taxonomy at scale across all 254 Clarivate Journal Citation Reports (JCR) (Clarivate, 2025) subject categories and 21 broader

disciplines. Conceptually, our taxonomy decomposes miscitation along five orthogonal dimensions—*status of the source, factual content, scope of application, evidence strength, and attribution link*—which jointly cover all failure modes observed in our corpus. MISCITEBENCH realizes this taxonomy as a *knowledge-blank, adversarial stress test* for LLM-based citation reasoning, rather than merely a labeled dataset. Table 1 summarizes how legacy labels from prior miscitation studies are absorbed into these categories.

New Category	Core Litmus Question	Absorbed Legacy Subtypes
Citation Validity Error	<i>Is the source itself disqualified from serving as scientific evidence?</i>	Obsolete or Retracted Citation; Secondary-Source Misuse
Content Misrepresentation Error	<i>Does the source, when read in context, actually say what the citing text claims it says?</i>	Irrelevant Citation; Contradictory Citation; Selective Quotation; Omitted Qualifiers; Conflated Findings
Scope Extrapolation Error	<i>Is an otherwise valid conclusion being applied outside the population, setting, or task for which it was established?</i>	Overgeneralization; Methodological Misapplication; Data-Scope Misuse
Evidence Characterization Error	<i>Is the type/strength of evidence claimed (e.g., causal, definitive) actually supported by the study design and statistics?</i>	Correlation-as-Causation; Statistical or Metrical Distortion
Attribution & Traceability Error	<i>Can a reader reliably locate and correctly attribute the source using the provided citation metadata?</i>	Ghost Citation; Author Misattribution

Table 1. Mapping from legacy miscitation labels to our unified 5-category taxonomy.

## 2.1. A Unified 5-Category Taxonomy of Miscitation

Starting from a survey of existing typologies and a manual audit of hundreds of real-world miscitations sampled across all 21 high-level disciplines, we consolidate previously scattered subtypes into five conceptually distinct categories. Each category is defined by a diagnostic litmus question that makes annotation operational and reproducible.

### 1. Citation Validity Error (Status of the Source).

The cited work itself lacks qualification as scientific evidence—for example, it has been retracted, superseded, or is a secondary source (e.g., a review or meta-analysis) being cited as if it were primary experimental

evidence. The error concerns the *status* of the source, not its content. Litmus question: “*Is the source itself disqualified from serving as scientific evidence?*” This category subsumes *Obsolete or Retracted Citation* and *Secondary-Source Misuse*.

### 2. Content Misrepresentation Error (Factual Content).

The citing text substantively distorts, fabricates, or reverses the findings, arguments, or conclusions of the source. Examples include citing a topically unrelated paper as evidence (*Irrelevant Citation*), citing a refutation as support (*Contradictory Citation*), or selectively quoting while omitting key qualifiers so that the meaning changes (*Selective Quotation, Omitted Qualifiers, Conflated Findings*). Litmus question: “*If I read the source in context, does it actually say what the citing sentence claims it says?*”

### 3. Scope Extrapolation Error (Scope of Application).

The source is correctly understood, but its conclusion is applied beyond the populations, settings, tasks, or methods for which it was established. Typical cases include *Overgeneralization* from narrow samples to broad populations, *Methodological Misapplication* outside validated constraints, and *Data-Scope Misuse* that promotes a subset analysis to a claim about the full dataset. Litmus question: “*Is an otherwise valid conclusion being applied outside the population, setting, or task for which it was established?*”

### 4. Evidence Characterization Error (Evidence Strength).

The citing text mischaracterizes the logical type, strength, or certainty of the evidence in the source. This includes treating correlational findings as causal (*Correlation-as-Causation*) or exaggerating statistical evidence (*Statistical or Metrical Distortion*), such as describing marginal effects as “conclusive causal evidence.” Litmus question: “*Is the type and strength of evidence claimed in the citing text actually supported by the source’s study design and statistics?*”

### 5. Attribution & Traceability Error (Attribution Link).

Errors in citation metadata break the link between claim and source. Examples include non-existent or unresolvable references (*Ghost Citation*) and assigning a result to the wrong author or paper (*Author Misattribution*). The error concerns the citation as a scholarly signpost, not the underlying evidence. Litmus question: “*Could a reader, using only this citation metadata, reliably locate the correct source and author of the claimed idea?*”

These five categories were sufficient to label all 6,350 miscitation instances in MISCITEBENCH (Section 2.2) without an “other” bucket. Annotators assign exactly one primary

category per instance, guided by the litmus questions above. When multiple error types co-occur, we enforce a logical *Dependency Precedence Rule* that mirrors the verification process: checking a citation aborts at the first point of failure. The precedence order is:

$$\begin{aligned} \text{Attribution \& Traceability} &\rightarrow \text{Citation Validity} \rightarrow \text{Content} \\ \text{Misrepresentation} &\rightarrow \text{Scope Extrapolation} \rightarrow \text{Evidence} \\ &\quad \text{Characterization.} \end{aligned}$$

For example, if the citation metadata is unusable (*Attribution*), one cannot assess retraction status (*Validity*) or content fidelity (*Content*), so Attribution becomes the primary label.

In an expert annotation study (Appendix B) where annotators could also choose “other” and “uncertain,” this protocol yielded Cohen’s  $\kappa$  in the “substantial” range, and fallback options were rarely used. Together with the absence of an “other” bucket in MISCITEBENCH, this provides empirical evidence that the taxonomy is both complete for our corpus and operationally consistent. In the remainder of this work, it serves as (i) the *label space* for MISCITEBENCH and (ii) the *error code space* predicted by our BIBAGENT.

## 2.2. MISCITEBENCH: A Contamination-Controlled Evaluation Framework

Existing misrepresentation datasets are typically narrow in domain, dominated by trivial errors, and vulnerable to data contamination: large language models may succeed by parametric memorization rather than by reasoning over the documents presented at evaluation time. We introduce **MISCITEBENCH**, a large-scale benchmark of 6,350 expert-validated misrepresentation instances that (i) spans all 254 Clarivate JCR subject categories and 21 high-level disciplines (including Agricultural Sciences, Clinical Medicine, Computer Science, Social Sciences, and Visual & Performing Arts), and (ii) is aligned with the five-category taxonomy above. MISCITEBENCH is constructed under two design principles: a *Knowledge-Blank Cleanroom Protocol* that filters out contaminated sources, and a *Dual-Tier Adversarial Generation* pipeline that yields both surface-level and deep-semantic misrepresentations.

**Source Selection and Knowledge-Blank Cleanroom Protocol.** For each of the 254 JCR subject categories, we identify the 2024 Journal Impact Factor (JIF) leader and then select, within that journal, the most-cited article published in 2024–2025; a detailed rationale for using most-cited articles as benchmark sources is provided in Appendix A.2. We manually retrieve the full text of each candidate source paper. To decouple reasoning from memorization, we probe a panel of frontier LLMs/LRMs (e.g., gpt-4o (Hurst et al., 2024), o4-mini-2025-04-06 (OpenAI, 2025b), claude-sonnet-4 (Anthropic, 2025); details in Appendix A.2) with 10 forensic questions per paper. Each ques-

tion is designed so that answering it correctly requires access to the main body or appendices (e.g., specific numerical results, methodological caveats, cross-section comparisons) and cannot be solved from title, abstract, or bibliographic metadata alone.

A candidate source is admitted only if every model in the panel answers *zero* of the  $N$  probes correctly. If any model answers at least one probe correctly, we discard that paper and test the next most-cited article in the same journal and time window, repeating until we obtain a source that satisfies this knowledge-blank criterion. This protocol substantially reduces the chance that models can rely on parametric knowledge of the paper’s content; at evaluation time they must instead reason over the contextual documents we provide. The procedure is model-agnostic and can be re-applied as stronger LLMs emerge.

**Dual-Tier Adversarial Generation and Expert Validation.** For each retained source paper, we use a state-of-the-art Large Reasoning Model (LRM; gemini-2.5-pro (Comanici et al., 2025; Google, 2025)) to generate 25 adversarial misrepresentations grounded in the taxonomy: 5 per category. Prompts expose the five category definitions and litmus questions so that each synthesized misrepresentation corresponds to a well-defined error code rather than an ad hoc negative example. Within each category, we instantiate a two-tier difficulty structure:

- **Surface-Level Misrepresentations** (3 per category): errors falsifiable by inspecting a single sentence or local paragraph, such as obviously irrelevant citations or explicit statistical distortion.
- **Deep-Semantic Misrepresentations** (2 per category): expert-level traps that mimic plausible scientific discourse but subtly violate global document logic, such as extrapolating conclusions to incompatible populations or conflating secondary and primary evidence in ways that require integrating results, limitations, and discussion.

Each instance packages (i) the erroneous citing sentence, (ii) the gold supporting span from the source, (iii) a natural-language explanation of the misrepresentation and its taxonomy label, and (iv) a corrected version of the citation. We then perform carefully and thoroughly cross-validation using an *independent* LRM (gpt-5.1-thinking (OpenAI, 2025a)) and human experts in the corresponding sub-discipline (verification criteria and prompts in Appendix A.3). Instances for which the two validators disagree are revised or discarded.

This pipeline yields a contamination-controlled benchmark with a consistent structure of 254 source papers  $\times$  5 taxonomy categories  $\times$  5 instances per category, for a total of 6,350 misrepresentation cases. MISCITEBENCH thus spans a wide range of di

MISCITEBENCH thus provides a spectrum of difficulty that includes both straightforward sanity-check failures and subtle expert-level miscitations, and—as shown in Section 4—serves as a stress test for miscitation detectors under distribution shift, especially for LLM agents that must reason over long, previously unseen scientific documents.

### 3. BIBAGENT: The Bibliographic Miscitation Detection Agent

We propose BIBAGENT, an end-to-end agentic framework designed to restore *traceable, citation-level accountability* to scientific discourse. Rather than processing pairs of papers monolithically in a single long LLM call, BIBAGENT orchestrates a modular pipeline that (i) handles both *accessible* and *inaccessible* (paywalled) sources and (ii) outputs miscitation judgements aligned with our five-category taxonomy, together with explicit citing contexts, evidence spans, and confidence scores. The system comprises four modules: (1) Document Parser & Citation Mapper (DPCM), (2) Cited Source Accessibility Classifier (CSAC), (3) Accessible Cited Source Verifier (ACSV), and (4) Inaccessible Cited Source Verifier (ICSV). Together, these modules take a single citing paper and produce both fine-grained, citation-level judgements and paper-level summaries of its overall citation integrity (Section 4).

#### 3.1. Citation Parsing and Accessibility Routing

The pipeline begins with the **Document Parser and Citation Mapper (DPCM)**, which accepts L<sup>A</sup>T<sub>E</sub>X source, XML/HTML, and PDF inputs and explicitly prioritizes *citation fidelity*. Crucially, DPCM normalizes every input—regardless of its original format—into a *structured Markdown intermediate representation* that preserves the document’s discourse skeleton: hierarchical headings, paragraph boundaries, inline math and displayed equations, figure/table captions, footnotes, and in-text citation anchors. This unified representation ensures that downstream verification operates on a consistent, traceable substrate rather than brittle format-specific text dumps.

For markup formats, DPCM removes non-semantic macros and formatting commands while preserving structural signals (e.g., section titles and their levels), mathematical expressions, and citation commands (e.g., `\cite`, `\ref`). It then renders the cleaned content into hierarchical Markdown with explicit heading levels and stable citation anchors, so that the logical argument flow and citation contexts remain intact across publishers and authoring styles.

For PDFs, DPCM uses a hybrid visual–linguistic parsing strategy that directly transcribes pages into the same structured Markdown representation. Each page is segmented into spatially coherent blocks via a sliding window, raster-

ized, and passed to a layout-aware multimodal model (i.e., gpt-4o-2024-08-06) together with any available extracted text. Guided by visual cues such as whitespace, font weight, column boundaries, and figure/caption geometry, the model performs *layout-grounded serialization*: it reconstructs the reading order of multi-column text and faithfully places floating figures, captions, sidebars, and footnotes into the appropriate Markdown locations, preserving both heading hierarchy and citation anchors. The exact multimodal prompting template and the image-to-Markdown transcription logic (including block serialization rules and failure-handling heuristics) are provided in Appendix C.

To minimize information loss, an *Extraction Verifier* audits the transcribed Markdown for structural and bibliographic continuity markers—monotonic section-heading progression, equation numbering, and citation index sequences (e.g., [12]→[13]). When it detects a discontinuity (e.g., a missing citation index, a broken equation sequence, or an implausible heading jump), it triggers localized re-parsing of the corresponding visual block at an adjusted resolution and re-integrates the repaired span back into the Markdown representation. This closed-loop verification mitigates OCR dropouts, misrecognized symbols, and fragmented sentences *before* downstream citation reasoning, ensuring that every later decision remains traceable to a faithful, structurally aligned document rendering.

The *Citation Mapper* then identifies in-text citation spans at the sentence level and links them to bibliographic entries, supporting more than 15 citation styles (e.g., APA, IEEE), grouped citations (e.g., “[3,5,9]”), cross-referenced footnotes, and idiosyncratic delimiters. The result is a structured mapping from localized citing contexts to their referenced sources, with style-normalized metadata for each link. All downstream judgements consume this mapping, which keeps every decision traceable to concrete citation contexts and bibliography entries. By committing all inputs to a single hierarchical Markdown representation with stable citation anchors, DPCM makes every downstream judgement auditable: each verdict can be traced back to the exact citing context and the exact source span in a format-invariant way.

The **Cited Source Accessibility Classifier (CSAC)** then routes each cited source. For every bibliographic entry, CSAC first attempts to resolve a DOI; if none is available, it constructs a metadata query from title, authors, and year. It queries official publisher or venue APIs to retrieve full text where possible. If this fails, it performs a secondary search over curated open-access repositories (e.g., arXiv, PubMed Central, SciELO, domain-specific preprint or institutional repositories). For each candidate match, CSAC compares title, author list, abstract, and (when available) reference list, accepting only open-access surrogates that are substantively identical to the publisher’s version.

CSAC also supports attribution checking. If the primary search yields no valid metadata record (e.g., invalid DOI) and the secondary repository search returns no high-similarity matches, the reference is flagged as a **Ghost Citation** (Attribution & Traceability Error) and directly finalized, preventing hallucinated rationales for non-existent papers. References with verified full text are routed to the *Accessible* stream. References with valid metadata but no full-text access are assigned a rich metadata snapshot (title, authors, abstract, venue, partial references) and routed to the *Inaccessible* stream, which is later consumed by ICSV.

### 3.2. ACSV: Adaptive Multi-Stage Verification

For accessible sources, the main challenge is to balance cost and reasoning depth. Naively feeding two full papers into a long-context LLM/LRM is expensive and prone to “lost-in-the-middle” errors. We instead use an **Adaptive Multi-Stage Verification** architecture that acts as a computational funnel: low-cost dense retrieval and NLI resolve easy cases, while only genuinely ambiguous citations are escalated to expensive deep reasoning. Empirically, this design underpins the MISCITEBENCH Open-regime gains reported in Section 4: across backbones, it improves miscitation detection accuracy over Full-Text baselines while reducing token usage by up to 79.4%.

**Phases I-II: Coarse Retrieval and Focused Re-ranking.** Let  $S_{\text{cite}}$  denote the citing sentence and  $D_{\text{cited}}$  the source document. We segment  $D_{\text{cited}}$  into paragraphs  $\mathcal{P} = \{p_i\}_{i=1}^M$ . A Bi-Encoder (instantiated as all-MiniLM-L6-v2 (Reimers & Gurevych, 2019)) maps  $S_{\text{cite}}$  and each  $p_i$  into a shared vector space and retrieves top- $K$  paragraphs via cosine similarity:

$$\text{Score}_{\text{retrieval}}(S_{\text{cite}}, p_i) = \cos(\mathbf{v}_{S_{\text{cite}}}, \mathbf{v}_{p_i}) = \frac{\mathbf{v}_{S_{\text{cite}}} \cdot \mathbf{v}_{p_i}}{\|\mathbf{v}_{S_{\text{cite}}}\| \|\mathbf{v}_{p_i}\|}. \quad (1)$$

A Cross-Encoder (e.g., ms-marco-BERT-base-v2 (Reimers & Gurevych, 2019)) then re-ranks the top- $K$  candidates and selects the top- $N$  segments  $\mathcal{P}_{\text{focus}}$ . In our experiments we set  $K = 10$  and  $N = 3$ , which empirically recovers sufficient context for complex arguments while keeping the evidence pool within the effective context window of downstream models.

To preserve inter-sentence dependencies while avoiding paragraph-level noise, we apply a sliding window over each paragraph in  $\mathcal{P}_{\text{focus}}$ , generating cited-side context windows  $\mathcal{W}_{\text{cited}}$  of  $W_{\text{size}}$  consecutive sentences with stride 1 (default  $W_{\text{size}} = 3$ ).

**Phase III: NLI-Based Logic Filtering with Dynamic Expansion.** At this stage, the goal is to decide, as cheaply as possible, whether the retrieved evidence already settles the citation. We therefore apply a Natural Language Inference

(NLI) model to each evidence window  $w \in \mathcal{W}_{\text{cited}}$ , treating the window as the *premise* and the citing sentence as the *hypothesis*. For every  $w$ , the model outputs a three-way distribution over *Entailment* ( $E$ ), *Contradiction* ( $C$ ), and *Neutral* ( $N$ ):

$$P(E, N, C | w, S_{\text{cite}}) = \text{NLI}(\text{premise} = w, \text{hypothesis} = S_{\text{cite}}). \quad (2)$$

We then implement an *Early Exit* rule that short-circuits easy cases. Let

$$M_E = \max_{w \in \mathcal{W}_{\text{cited}}} P(E | w, S_{\text{cite}}), \quad (3)$$

$$M_C = \max_{w \in \mathcal{W}_{\text{cited}}} P(C | w, S_{\text{cite}}), \quad (4)$$

denote the strongest entailment and contradiction signals across all windows. Whenever either signal crosses a high-confidence threshold  $\tau_{\text{high}}$ , we immediately commit to a decision:

$$\text{Decision} = \begin{cases} \text{Correct}, & M_E > \tau_{\text{high}}, \\ \text{Miscitation}, & M_C > \tau_{\text{high}}. \end{cases} \quad (5)$$

When neither entailment nor contradiction exceeds  $\tau_{\text{high}}$ , or when both do so with conflicting labels, the case is treated as *ambiguous*. To address this, we dynamically expand the hypothesis by incorporating its immediate neighbors in the citing document:

$$S_{\text{expanded}} = \text{concat}(S_{\text{prev}}, S_{\text{cite}}, S_{\text{next}}), \quad (6)$$

and re-run NLI with  $S_{\text{expanded}}$  as the hypothesis while keeping  $\mathcal{W}_{\text{cited}}$  as the set of premises. In all experiments, we instantiate the NLI model with a publicly available DeBERTa-v3-large checkpoint that is pre-trained for NLI, use it *off-the-shelf* without any additional fine-tuning, and fix  $\tau_{\text{high}} = 0.9$ .

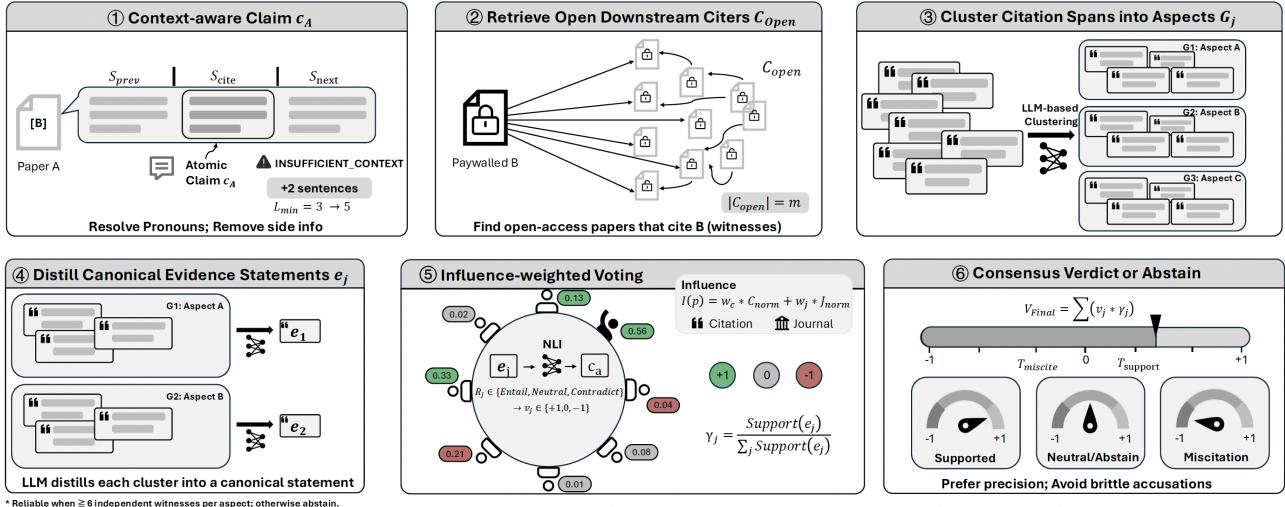
**Phase IV: LRM Deep Reasoning via Self-Consistency.** Cases that remain ambiguous enter the most expensive phase. We construct a Chain-of-Thought prompt containing  $S_{\text{expanded}}$  and  $\mathcal{P}_{\text{focus}}$  and query a Large Reasoning Model (LRM, e.g., gemini-2.5-pro) to perform semantic arbitration (supported vs. miscitation vs. undecidable). To mitigate stochastic artifacts, we enforce *Self-Consistency*: we query the LRM  $M$  times with different sampling seeds at temperature  $T$  (default  $M = 5$ ,  $T = 0.7$ ), obtaining verdicts  $\{V_1, \dots, V_M\}$ , and adopt the majority class as the final label. We define a confidence score

$$\text{Confidence}_{\text{LRM}} = \frac{\text{Count}(\text{Majority Verdict})}{M}. \quad (7)$$

If  $\text{Confidence}_{\text{LRM}} < 0.6$ , ACSV outputs an “Undecidable—Requires manual review” label instead of forcing a brittle decision.

### Paywalled Source B is Verified via Community Consensus – Not By Hallucinating B.

aggregate downstream open-access witnesses  $\rightarrow$  cluster spans  $\rightarrow$  distill evidence statements  $\rightarrow$  influence-weighted voting  $\rightarrow$  reliable consensus or abstain



**Figure 1.** Overview of the INACCESSIBLE CITED SOURCE VERIFIER (ICSV) and its Evidence Committee mechanism. Given a citing context about a paywalled source  $B$ , ICSV (1) extracts an atomic claim that captures exactly what the citing paper attributes to  $B$ ; (2) retrieves open-access downstream citers of  $B$  and clusters their local citation contexts into aspect-specific groups; (3) distills each group into a canonical evidence statement, weighting it by a field-normalized influence score that combines venue and paper-level impact; and (4) aggregates the resulting entailment/contradiction/neutral votes into a reliability-aware consensus verdict, explicitly abstaining when community evidence is too sparse or internally inconsistent. This converts paywalled miscitation detection from an inaccessible-document problem into a traceable, community-consensus reasoning task.

### 3.3. ICSV: The Evidence Committee Mechanism

The **Inaccessible Cited Source Verifier (ICSV)** handles sources  $B$  whose full text is behind a paywall or otherwise inaccessible. Figure 1 summarizes the resulting Evidence Committee pipeline that reconstructs the community’s view of such sources. Since direct textual entailment against  $B$  is impossible, ICSV adopts a *Community Consensus Reconstruction* approach: it treats downstream citations as a distributed, noisy memory of  $B$ ’s contributions and reconstructs those contributions by aggregating and weighting statements from open-access witness papers. Individual downstream citers may miscite  $B$ , so ICSV explicitly models each evidence statement as a noisy vote and relies on field-normalized influence weights and abstention thresholds to maintain reliability. Throughout this module, all LLM calls share a single backbone (gpt-4o-2024-08-06) but are driven by task-specific prompt templates and decoding hyperparameters; we document these prompts, sampling configurations, and post-processing rules in detail in Appendix D.

**(1) Context-Aware Citing Claim Extraction.** Given a citing sentence  $s_A$  in paper  $A$  that references inaccessible source  $B$ , we construct a local context window  $W_A$  by concatenating the preceding, current, and succeeding sentences (default  $L_{\text{win}} = 3$ ). An LLM is prompted to extract from  $W_A$  a single, self-contained atomic claim  $c_A$  that captures

exactly what  $A$  attributes to  $B$ , resolving pronouns and removing side information. The prompt admits a special INSUFFICIENT\_CONTEXT token; if returned, we expand  $W_A$  (e.g., to  $L_{\text{win}} = 5$ ) and retry until we obtain a stable  $c_A$  or mark the case underspecified.

### (2) Committee Formation and Evidence Distillation.

We retrieve a set of open-access downstream citers  $C_{\text{open}} = \{p_1, \dots, p_m\}$  that reference  $B$ . For each mention of  $B$ , we extract a context-rich span comprising the citing sentence and its neighbors, forming a set  $S = \{s_1, \dots\}$ . We embed all spans and cluster them; an LLM is used to refine clusters so that each  $G_j$  represents a coherent aspect of  $B$  (e.g., algorithmic contribution vs. dataset construction). For each cluster  $G_j$ , an LLM distills its spans into a canonical *Evidence Statement*  $e_j$ , yielding an evidence set  $E = \{e_1, \dots, e_k\}$  that summarizes the community’s distributed view of  $B$  along multiple semantic axes.

**(3) Field-Normalized Influence Modeling.** Citation and venue statistics vary drastically across fields and years. We therefore define a **Field-Normalized Influence Score**  $I(p)$  for each committee member  $p \in C_{\text{open}}$ . Let  $\text{IF}(p)$  be the impact factor of  $p$ ’s venue and  $\text{Cite}(p)$  its raw citation count. We normalize both against the appropriate JCR Subject

Category and publication year:

$$J_{\text{norm}}(p) = \text{Rank}\%(\text{IF}(p) | \text{Field}(p)), \quad (8)$$

$$C_{\text{norm}}(p) = \text{Rank}\%(\text{Cite}(p) | \text{Field}(p), \text{Year}(p)), \quad (9)$$

and aggregate them as

$$\mathcal{I}(p) = w_c \cdot C_{\text{norm}}(p) + w_j \cdot J_{\text{norm}}(p), \quad (10)$$

where we set  $w_c = 0.6$  and  $w_j = 0.4$ . Crucially, we assign higher weight to paper-level citation performance ( $w_c$ ) than venue prestige ( $w_j$ ). This design choice explicitly mitigates the “halo effect” of top-tier journals, ensuring that highly influential papers published in niche or lower-impact venues are correctly recognized as credible evidence sources by the committee.

The credibility of an evidence statement  $e_j$  is the normalized sum of influences from all papers contributing spans to cluster  $G_j$ . Let  $\text{SourcePaper}(s)$  denote the paper from which span  $s$  was extracted. We define

$$\text{Support}(e_j) = \sum_{s \in G_j} \mathcal{I}(\text{SourcePaper}(s)), \quad (11)$$

$$\gamma_j = \frac{\text{Support}(e_j)}{\sum_{i=1}^k \text{Support}(e_i)}, \quad (12)$$

where  $\gamma_j \in [0, 1]$  is the credibility weight of  $e_j$ .

#### (4) Reliability-Aware Weighted Consensus Verdict.

For each  $e_j$ , an LLM classifies the relation  $R_j \in \{\text{Entailment}, \text{Contradiction}, \text{Neutral}\}$  between  $e_j$  and  $c_A$ , which we map to a scalar vote  $v_j \in \{+1, 0, -1\}$ . We then compute a credibility-weighted consensus score

$$\mathcal{V}_{\text{final}} = \sum_{j=1}^k v_j \cdot \gamma_j, \quad (13)$$

with  $\mathcal{V}_{\text{final}} \in [-1, 1]$ . A citation is labeled *Supported* if  $\mathcal{V}_{\text{final}} > T_{\text{support}}$ , *Miscitation* if  $\mathcal{V}_{\text{final}} < T_{\text{miscite}}$ , and *Undecidable* otherwise, with  $|\mathcal{V}_{\text{final}}|$  serving as a confidence score. In all experiments we set  $T_{\text{support}} = 0.3$  and  $T_{\text{miscite}} = -0.3$ .

Community evidence is sparse when few downstream citers exist. Our pilot studies (Section 4) show a sharp increase in verdict stability once an aspect is supported by at least  $K_{\min} = 6$  independent witnesses. Guided by this, ICSV enforces a *Reliability-Aware Abstention* protocol: if  $|C_{\text{open}}| < K_{\min}$  or if  $\mathcal{V}_{\text{final}} \in [T_{\text{miscite}}, T_{\text{support}}]$ , ICSV abstains instead of forcing a brittle verdict. This design prioritizes precision over recall in high-stakes settings, avoiding false accusations when community evidence is weak while exploiting the research community’s distributed memory when it is sufficiently rich.

### 3.4. Taxonomy-Aligned Labeling

Beyond deciding whether a citation is factually valid, BIBAGENT aims to assign a fine-grained miscitation type from the five-category taxonomy. For every citation flagged as a miscitation by ACSV or ICSV, we invoke a lightweight taxonomy classifier that operates over the same evidence context, augmented with CSAC-derived metadata (e.g., article type, retraction status) where available.

To enforce the *Dependency Precedence Rule* from Section 2.1, the classifier is prompted with a hierarchical decision tree: it must first check for *Attribution & Traceability* failures using CSAC’s resolution results; if none apply, it considers *Citation Validity* (e.g., retraction, secondary-source misuse), then *Content Misrepresentation*, and only then *Scope Extrapolation* and *Evidence Characterization*. The input bundle comprises (i) the citing context  $S_{\text{expanded}}$ , (ii) key evidence windows from  $\mathcal{W}_{\text{cited}}$  (for ACSV) or distilled evidence statements  $\{e_j\}$  (for ICSV), and (iii) the five taxonomy definitions with their litmus questions. A compact LLM (i.e., gpt-4o-2024-08-06) is used in a zero-shot, small-ensemble self-consistency setting to choose exactly one category and provide a short rationale; we take the majority vote across runs.

In summary, for each citation in the input paper, BIBAGENT outputs (i) a validity judgement (*Supported*, *Miscitation*, or *Undecidable*), (ii) a single taxonomy-aligned error code for miscitations, (iii) the citing context with key evidence spans (or distilled evidence statements for inaccessible sources), and (iv) a scalar confidence score. These structured outputs underpin the paper-level citation integrity summaries reported in Section 4, and support high-recall miscitation detection with interpretable, taxonomy-grounded explanations at practical computational cost.

## 4. Experiments and Results

### 4.1. Evaluation Protocol and Deployment Regimes

We evaluate BIBAGENT on **MisciteBench** under two regimes that mirror the two dominant realities of citation verification: (i) the *full-text* regime, where verification should be grounded, efficient, and traceable, and (ii) the *paywall* regime, where the full text of the cited source is unavailable, and verification must remain reliable without fabricating what the unseen source contains. These regimes directly probe the two central claims of this work: BIBAGENT can (a) *compress* long-document verification without sacrificing diagnostic fidelity, and (b) *reconstruct* evidence for inaccessible sources via community consensus rather than brittle retrieval or speculation.

**Regime I: MisciteBench-Open (Accessible Sources).** In this setting, the full text of the cited source is available. Each

instance provides the citing sentence (with its local citing context) and the complete cited paper, and is routed to ACSV. The system outputs (i) a validity judgment (*Supported* vs. *Miscitation*) and (ii) a diagnosis that must be faithful to the gold miscitation rationale under our taxonomy.

**Regime II: MisciteBench-Paywall (Inaccessible Sources).** In this setting, the cited source text is *not* provided; only bibliographic metadata and downstream open-access citers are available, and the instance is routed to ICSV. This is a strict inaccessibility condition rather than a contrived ablation: MisciteBench is constructed via a *knowledge-blank cleanroom protocol* (Section 2.2) that filters out any source paper whose content can be answered correctly by the tested backbones through parametric memorization. As a result, withholding the source text genuinely forces models to operate under paywall constraints.

**Evaluation under Paywall Constraints.** MisciteBench includes both *surface-level* and *deep-semantic* miscitations. Surface-level citations refer to relatively straightforward factual references, such as key statistics, population characteristics, sample sizes, or other clearly stated descriptive elements, that can typically be verified without requiring a comprehensive reading or deep conceptual understanding of the cited article. In contrast, deep semantic miscitations involve claims that depend on a holistic interpretation of the cited work, often requiring integration across its results, assumptions, limitations, and discussion. When the full text of the cited paper is inaccessible, evaluating deep semantic miscitations becomes intrinsically underdetermined. Moreover, prior research on citation function has shown that many citations serve descriptive or perfunctory roles rather than deep substantive engagement (Budi & Yaniasih, 2023; Kunnath et al., 2021), suggesting that surface-level verification captures a substantial and practically relevant portion of how scholarly citations are actually used. Accordingly, under the paywalled setting, we restrict evaluation to the surface-level subset of MisciteBench, yielding 3 instances per category  $\times$  5 categories  $\times$  254 fields, for a total 3,810 examples.

## 4.2. Baselines and Why Backbone-Controlled Comparisons Are Necessary

Miscitation detection faces not only methodological heterogeneity but also a persistent *reproducibility and applicability gap* in existing benchmarks and baselines. Despite substantial work on citation sentiment and scientific claim verification (Wadden et al., 2022; 2020; Press et al., 2024; Qian et al., 2025), most available “miscitation” datasets and systems rely on small, domain-limited collections, short-context evaluation, or incompletely documented pipelines, making fair, end-to-end comparison difficult. In practice, data are often inaccessible, paywalled, or structurally in-

complete for long-context verification; links to evaluation code or checkpoints are missing; and many systems lack support for long-document evidence and remain tailored to abstract-level or snippet-level entailment, misaligned with MisciteBench’s long-document, diagnosis-critical setting. Under these conditions, reporting superficial cross-paper numbers risks creating only the illusion of comparability while failing the standards of reproducible science.

We therefore adopt **architecture-agnostic, backbone-controlled** baselines that isolate the incremental contribution of BIBAGENT’s agentic decomposition, while remaining fully reproducible and strictly matched in model capacity:

- **Full-Text** (Open regime): concatenate the citing context and the entire cited paper into a single prompt and ask the same backbone to decide *Supported* vs. *Miscitation* and explain why.
- **Search** (Paywall regime): allow the backbone to call web search tools to locate potential open-access surrogates and then verify against retrieved snippets or documents. This baseline is intentionally generous: it may bypass paywalls whenever surrogates exist.
- **BIBAGENT**: the same backbone embedded inside our pipeline (ACSV for Open; ICSV for Paywall). Crucially, ICSV runs *offline* without external search to measure whether BIBAGENT can preserve citation integrity when retrieval is fundamentally unreliable or incomplete.

This backbone-controlled design answers the core scientific question of this paper without confounding: *given the same LLM/LRM, does the agentic verification structure (retrieval/NLI funnel; committee consensus) convert long-context and paywall verification from a brittle prompt into a reliable procedure?* In other words, we treat the lack of reliable baselines not as an inconvenience but as an empirical fact about the field’s current tooling: a **community reproducibility gap** that motivates, rather than undermines, backbone-matched evaluation.

## 4.3. Metrics

**Primary metric: Acc-pass@3.** We measure both correctness and diagnostic faithfulness using **Acc-pass@3**. For each instance, a method is sampled three times with different decoding seeds. A prediction is counted as correct if *any* of the three outputs simultaneously:

- predicts the correct validity label (*Supported* vs. *Miscitation*), and
- provides an explanation that is *semantically equivalent*

to the gold miscitation rationale (i.e., it identifies the failure mechanism, not merely the label).

Semantic equivalence is judged by an independent LLM grader (`gpt-4o-2024-08-06`) under a rubric that penalizes partial matches, generic paraphrases, and post-hoc hallucinations (details and robustness checks in Appendix E). This metric reflects the editorial requirement that a detector must not only flag a miscitation but also justify it with traceable reasoning.

**Efficiency metric: Token Economy.** For MisciteBench-Open, we additionally report **Token Economy**—the relative reduction in total tokens processed per instance (inputs + outputs) compared to the corresponding Full-Text baseline using the same backbone, computed on instances where both methods return a verdict. Token accounting and decoding settings are described in Appendix E.

Model	Org.	Scenario	Acc-pass@3 ↑	Token Econ ↑
gpt-5-2025-08-07		Full-Text BIBAGENT	92.1 <b>98.8</b>	65.3
o4-mini-2025-04-16		Full-Text BIBAGENT	88.4 <b>96.4</b>	79.4
gpt-4o-2024-08-06		Full-Text BIBAGENT	84.6 <b>94.8</b>	60.2
gpt-oss-120b		Full-Text BIBAGENT	82.9 <b>92.7</b>	59.3
gpt-oss-20b		Full-Text BIBAGENT	72.4 <b>87.6</b>	70.0
claude-sonnet-4-20250514		Full-Text BIBAGENT	91.3 <b>97.8</b>	74.7
claude-opus-4-20250514		Full-Text BIBAGENT	94.3 <b>100.0</b>	44.6
gemini-2.5-pro		Full-Text BIBAGENT	95.8 <b>97.9</b>	74.8
gemini-3-flash		Full-Text BIBAGENT	93.1 <b>96.4</b>	66.2
gemini-3-pro		Full-Text BIBAGENT	96.8 <b>100.0</b>	64.7
Nemotron 3 Nano (30B A3B)		Full-Text BIBAGENT	73.0 <b>88.4</b>	72.1
Llama-3.3 Nemotron Super 49B v1		Full-Text BIBAGENT	80.8 <b>89.9</b>	66.4
Qwen3-235B-A22B-Thinking-2507		Full-Text BIBAGENT	86.4 <b>95.5</b>	61.8
Qwen3 VL 32B Thinking		Full-Text BIBAGENT	79.3 <b>92.4</b>	66.6
Qwen3 VL 8B Thinking		Full-Text BIBAGENT	54.8 <b>80.2</b>	76.5
Minstral 3 (14B Reasoning 2512)		Full-Text BIBAGENT	56.2 <b>82.1</b>	74.5
Magistral Medium		Full-Text BIBAGENT	66.4 <b>87.0</b>	68.3
Deepseek-V3.2 (Thinking)		Full-Text BIBAGENT	88.0 <b>98.6</b>	62.4
DeepSeek-R1-0528		Full-Text BIBAGENT	92.4 <b>97.2</b>	62.1
Deepseek R1 Distill Qwen 32B		Full-Text BIBAGENT	76.6 <b>92.0</b>	68.7
Deepseek R1 Distill Qwen 14B		Full-Text BIBAGENT	68.4 <b>88.1</b>	72.9
Llama 3.1 405B Instruct		Full-Text BIBAGENT	74.5 <b>91.0</b>	60.2
Llama 3.3 70B Instruct		Full-Text BIBAGENT	70.3 <b>84.6</b>	63.1

Table 2. MisciteBench-Open: miscitation detection and diagnosis when full texts of cited sources are accessible. Token Econ reports the percentage of tokens saved by BIBAGENT relative to the corresponding Full-Text baseline.

#### 4.4. Results on MisciteBench-Open (Accessible Sources)

For accessible references, we compare BIBAGENT to **Full-Text** baselines. Full-Text is the most common “obvious” strategy in practice, but it conflates two failure modes that matter for citation integrity: (i) *long-context brittleness* (models miss or dilute the decisive evidence in the middle of a full paper), and (ii) *explanation drift* (models generate fluent rationales that are weakly grounded in the cited text). ACSV explicitly targets both by enforcing a “zoom-in” verification funnel: retrieve a small, high-recall evidence pool; apply calibrated NLI with dynamic citing-context expansion; and escalate to LRM arbitration only when ambiguity persists.

Table 2 shows that ACSV improves both effectiveness and efficiency across all backbones. BIBAGENT yields consistent gains in Acc-pass@3 (from **+5.7** up to **+19.8** absolute points), with the largest gains appearing precisely where Full-Text prompting is most fragile under long context (e.g., `gpt-4o`). At the same time, ACSV reduces token usage by **44.6–79.4%**, confirming that the “adaptive zoom-in” architecture resolves most instances before expensive arbitration, without sacrificing diagnostic fidelity.

**Failure patterns (what Full-Text gets wrong).** Qualitatively, the dominant Full-Text failures are not random noise but systematic: errors concentrate on cases where the decisive evidence is (i) expressed across multiple neighboring sentences requiring local coherence, (ii) embedded in longer chains of scientific qualification (e.g., limitations or conditional claims), or (iii) easy to paraphrase fluently yet hard to ground precisely. In these regimes, monolithic prompting often produces confident but weakly supported explanations, while ACSV’s evidence funnel forces the decision to be anchored to a small, explicitly retrieved  $\mathcal{P}_{\text{focus}}$  and its derived windows, reducing both long-context dilution and rationale drift.

#### 4.5. Results on MisciteBench-Paywall (Inaccessible Sources)

Paywall verification is the setting where the field most often fails in practice. Here, models face a fundamental dilemma: either abstain (which does not scale to editorial needs) or speculate (which destroys traceability). We compare ICSV against the **Search** baseline, which is intentionally optimistic: it may retrieve open surrogates that bypass paywalls, and thus is given the best possible chance to succeed.

The gap is decisive. Even with web search enabled, Search baselines achieve only **22–36** Acc-pass@3 and frequently fail the diagnostic requirement: they may retrieve incomplete or mismatched versions, conflate unrelated snippets, or generate plausible-sounding rationales unsupported by verifiable evidence. In contrast, ICSV improves Acc-pass@3

Model	Org.	Scenario	Acc-pass@3 ↑
gpt-5-2025-08-07		Search BIBAGENT	36.2 <b>80.3</b>
o4-mini-2025-04-16		Search BIBAGENT	29.8 <b>66.4</b>
claude-sonnet-4-20250514		Search BIBAGENT	30.2 <b>66.8</b>
claude-opus-4-20250514		Search BIBAGENT	34.4 <b>74.2</b>
gpt-4o-2024-08-06		Search BIBAGENT	22.1 <b>66.5</b>
gpt-oss-120b		Search* BIBAGENT	16.4 <b>69.4</b>
gpt-oss-20b		Search* BIBAGENT	6.6 <b>55.7</b>
Nemotron 3 Nano (30B A3B)		Search* BIBAGENT	12.3 <b>56.2</b>
Llama-3.3 Nemotron Super 49B v1		Search* BIBAGENT	11.6 <b>59.3</b>
Qwen3-235B-A22B-Thinking-2507		Search* BIBAGENT	18.2 <b>69.2</b>
Qwen3 VL 32B Thinking		Search* BIBAGENT	13.7 <b>63.9</b>
Qwen3 VL 8B Thinking		Search* BIBAGENT	2.1 <b>54.2</b>
Minstral 3 (14B Reasoning 2512)		Search BIBAGENT	17.9 <b>56.3</b>
Magistral Medium		Search BIBAGENT	28.0 <b>62.4</b>
Deepseek-V3.2 (Thinking)		Search* BIBAGENT	14.8 <b>76.8</b>
DeepSeek-R1-0528		Search* BIBAGENT	14.5 <b>70.6</b>
Deepseek R1 Distill Qwen 32B		Search* BIBAGENT	9.6 <b>62.7</b>
Deepseek R1 Distill Qwen 14B		Search* BIBAGENT	8.2 <b>59.6</b>
Llama 3.1 405B Instruct		Search* BIBAGENT	10.3 <b>64.3</b>
Llama 3.3 70B Instruct		Search* BIBAGENT	7.1 <b>59.4</b>

Table 3. MsciteBench-Paywall: miscitation detection when the source text is inaccessible. “Search” and “Search\*” isolate two fundamentally different retrieval regimes: **Search** uses *API-native web search* that is inherited by the model at call time (OpenAI: GPT-5/o4-mini/GPT-4o; Anthropic: Sonnet 4/Opus 4; Mistral Agents: Minstral/Magistral), whereas **Search\*** uses our *self-implemented* search tool, invoked via tool calling for all other models (DeepSeek /NVIDIA /Qwen /Llama /gpt-oss). BIBAGENT uses only ICSV’s Evidence Committee without any external tools.

to **36.5–80.3** across backbones, more than doubling accuracy in every case and reaching **80.3** with gpt-5. This validates the core thesis of ICSV: paywalled miscitation detection is not a retrieval problem but a *reliability problem*. The Evidence Committee mechanism solves this by (i) constructing multiple independent witness views, (ii) distilling them into coherent evidence statements, and (iii) aggregating them with field-normalized credibility weights and reliability-aware abstention.

**Failure patterns (what Search gets wrong).** The dominant Search failure mode is structural: even when retrieval succeeds, the evidence is often fragmented, version-

mismatched, or context-stripped, making the model prone to selecting the first plausible snippet and then reverse-engineering an explanation. When retrieval fails (or retrieves the wrong source), the baseline degenerates into either abstention or hallucinated reconstruction. In contrast, ICSV never claims to “read” the paywalled paper; it instead reconstructs what the community collectively attributes to it, and explicitly abstains when the witness set is too small or too contradictory.

#### 4.6. When Does the Evidence Committee Become Reliable?

A practical question for deployment is how much community evidence is needed for stable paywall verification. We perform an ablation over the cluster size  $|G_j|$  for each evidence statement  $e_j$  (Appendix F). The results reveal a sharp reliability transition: when an evidence statement is supported by more than **6** independent downstream witnesses ( $|G_j| > 6$ ), BIBAGENT typically produces high-confidence verdicts (confidence  $\geq 90\%$ ), while below this threshold the framework abstains more frequently rather than forcing brittle decisions. This behavior is a deliberate integrity constraint: in high-stakes workflows, an abstention with an explicit evidence shortage is preferable to an overconfident accusation grounded in weak or noisy community memory.

**Summary.** Across both regimes, BIBAGENT converts miscitation detection from a monolithic prompting heuristic into a traceable verification procedure. On accessible sources, its adaptive zoom-in funnel improves diagnostic accuracy while the cutting token cost by up to **79.4%**. On paywalled sources, its Evidence Committee more than doubles accuracy compared to a search-augmented baseline, despite operating offline and without ever claiming access to the primary record. Together, these results establish a concrete path to making citation chains auditable at scale in the GenAI era: *efficient when evidence is available, and principled when it is not*.

## 5. Conclusion

This work takes a step toward making scientific citation chains *auditable by default*. Starting from a field fragmented by ad hoc error labels, short-context datasets, and paywall limitations, we introduce three components that together reframe miscitation detection as an end-to-end, traceable reasoning problem. First, our five-category taxonomy provides an operational error code space that is both mutually exclusive and collectively exhaustive across 254 sub-disciplines, turning vague intuitions about “bad citations” into a structured object of study. Second, MsciteBench offers a contamination-controlled, knowledge-blank benchmark of 6,350 instances that stress-tests models on both

surface-level and expert-level miscitations, rather than rewarding memorization of a few canonical sources. Third, BibAgent itself demonstrates that miscitation can be detected and *explained* with high reliability by an agentic pipeline that integrates symbolic routing, efficient retrieval, calibrated NLI, and committee-based reasoning.

Empirically, BibAgent closes two longstanding deployment gaps. On accessible sources, its adaptive “zoom-in” architecture matches or surpasses full-document LLM baselines while cutting token usage by up to **79.4%**, turning long-document verification from a prohibitively expensive operation into a tractable one. On inaccessible sources, its Evidence Committee mechanism more than doubles miscitation detection accuracy compared to web-search-augmented LLMs, despite operating without any direct access to the paywalled text. In an era where generative models can produce polished but potentially ungrounded manuscripts at scale, this combination—a principled taxonomy, a hard benchmark, and a paywall-robust agent—offers a concrete path toward self-auditing scientific workflows: draft with one model, verify with another, and keep the citation graph honest.

Looking ahead, BibAgent opens several avenues for building *self-correcting* research ecosystems: tight integration with authoring tools and journal submission pipelines; extensions to multilingual literature and non-textual evidence (e.g., figures and code); and defenses against adversarially crafted miscitations designed to exploit model biases. But the central message is already clear. Mispuration is no longer an unavoidable side effect of scale; with the right abstractions and reasoning architecture, it becomes a measurable, diagnosable, and ultimately controllable property of scientific communication.

Finally, while our current focus remains on scholarly publications, the agentic architecture of BibAgent is fundamentally extensible. Its ability to reconstruct evidence through community consensus offers a transformative blueprint for auditing other high-stakes documents, such as grant proposals, patent applications, and policy briefs, ensuring that critical societal decisions are anchored in accurate and contextually faithful evidence. Looking ahead, we envision BibAgent enabling a broader culture of verifiable knowledge, promoting transparency, accountability, and trust across domains where evidence integrity is paramount.

## References

- Anthropic. Introducing claude 4, 2025. URL <https://www.anthropic.com/news/clause-4>.
- Athar, A. and Teufel, S. Detection of implicit citations for sentiment detection. In Van Den Bosch, A. and Shatkay, H. (eds.), *Proceedings of the Workshop on Detecting Structure in Scholarly Discourse*, pp. 18–26, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL <https://aclanthology.org/W12-4303/>.
- Budi, I. and Yaniasih, Y. Understanding the meanings of citations using sentiment, role, and citation function classifications. *Scientometrics*, 128(1):735–759, 2023.
- Chawla, D. S. The citation black market: schemes selling fake references alarm scientists. *Nature*, 632(8027):966–966, 2024.
- Clarivate. Journal citation reports, 2025. URL <https://clarivate.com/academia-government/scientific-and-academic-research/research-funding-analytics/journal-citation-reports/>.
- Comanici, G., Bieber, E., Schaeckermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.
- Dai, C., Chen, Q., Wan, T., Liu, F., Gong, Y., and Wang, Q. Literary runaway: Increasingly more references cited per academic research article from 1980 to 2019. *Plos one*, 16(8):e0255849, 2021.
- De Lacey, G., Record, C., and Wade, J. How accurate are quotations and references in medical journals? *Br Med J (Clin Res Ed)*, 291(6499):884–886, 1985.
- Google. Gemini 2.5 pro model card, 2025. URL <https://modelcards.withgoogle.com/assets/documents/gemini-2.5-pro.pdf>. Accessed: 2026-01-11.
- Greenberg, S. A. How citation distortions create unfounded authority: analysis of a citation network. *Bmj*, 339, 2009.
- Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.
- Jergas, H. and Baethge, C. Quotation accuracy in medical journal articles—a systematic review and meta-analysis. *PeerJ*, 3:e1364, 2015.
- Kunnath, S. N., Herrmannova, D., Pride, D., and Knoth, P. A meta-analysis of semantic classification of citations. *Quantitative science studies*, 2(4):1170–1215, 2021.
- Liang, W., Zhang, Y., Wu, Z., Lepp, H., Ji, W., Zhao, X., Cao, H., Liu, S., He, S., Huang, Z., et al. Mapping the

- increasing use of llms in scientific papers. *arXiv preprint arXiv:2404.01268*, 2024.
- OpenAI. Gpt-5.1: A smarter, more conversational chatgpt, 2025a. URL <https://openai.com/index/gpt-5-1/>.
- OpenAI. Introducing openai o3 and o4-mini, 2025b. URL <https://openai.com/index/introducing-o3-and-o4-mini/>.
- Press, O., Hochlehnert, A., Prabhu, A., Udandarao, V., Press, O., and Bethge, M. Citeme: Can language models accurately cite scientific claims?, 2024. URL <https://arxiv.org/abs/2407.12861>.
- Pride, D. and Knoth, P. Incidental or influential?-challenges in automatically detecting citation importance using publication full texts. In *International conference on theory and practice of digital Libraries*, pp. 572–578. Springer, 2017.
- Pride, D. and Knoth, P. An authoritative approach to citation classification. In *Proceedings of the ACM/IEEE joint conference on digital libraries in 2020*, pp. 337–340, 2020.
- Qian, H., Fan, Y., Guo, J., Zhang, R., Chen, Q., Yin, D., and Cheng, X. Vericite: Towards reliable citations in retrieval-augmented generation via rigorous verification. In *Proceedings of the 2025 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, pp. 47–54. ACM, December 2025. doi: 10.1145/3767695.3769505. URL <http://dx.doi.org/10.1145/3767695.3769505>.
- Rao, V. S., Kumar, A., Lakkaraju, H., and Shah, N. B. Detecting llm-generated peer reviews. *PLoS One*, 20(9):e0331871, 2025.
- Reimers, N. and Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL <http://arxiv.org/abs/1908.10084>.
- Rekdal, O. B. Academic urban legends. *Social Studies of Science*, 44(4):638–654, 2014.
- Sarol, M. J., Ming, S., Radhakrishna, S., Schneider, J., and Kilicoglu, H. Assessing citation integrity in biomedical publications: corpus annotation and nlp models. *Bioinformatics*, 40(7):btae420, 2024.
- Secchi, D. A simple model of citation cartels: when self-interest strikes science. In *Conference of the European Social Simulation Association*, pp. 23–32. Springer, 2022.
- Simkin, M. V. and Roychowdhury, V. P. Copied citations create renowned papers? *arXiv preprint cond-mat/0305150*, 2003.
- Teufel, S., Siddharthan, A., and Tidhar, D. Automatic classification of citation function. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pp. 103–110, 2006.
- Wadden, D. and Lo, K. Overview and insights from the sciver shared task on scientific claim verification. *arXiv preprint arXiv:2107.08188*, 2021.
- Wadden, D., Lin, S., Lo, K., Wang, L. L., van Zuylen, M., Cohan, A., and Hajishirzi, H. Fact or fiction: Verifying scientific claims. *arXiv preprint arXiv:2004.14974*, 2020.
- Wadden, D., Lo, K., Wang, L. L., Cohan, A., Beltagy, I., and Hajishirzi, H. MultiVerS: Improving scientific claim verification with weak supervision and full-document context. In Carpuat, M., de Marneffe, M.-C., and Meza Ruiz, I. V. (eds.), *Findings of the Association for Computational Linguistics: NAACL 2022*, pp. 61–76, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.6. URL <https://aclanthology.org/2022.findings-naacl.6>.
- Walters, W. H. and Wilder, E. I. Fabrication and errors in the bibliographic citations generated by chatgpt. *Scientific Reports*, 13(1):14045, 2023.
- Waltman, L. A review of the literature on citation impact indicators. *Journal of informetrics*, 10(2):365–391, 2016.
- Wei, J., Huang, D., Lu, Y., Zhou, D., and Le, Q. V. Simple synthetic data reduces sycophancy in large language models, 2024. URL <https://arxiv.org/abs/2308.03958>.
- Wilhite, A. W. and Fong, E. A. Coercive citation in academic publishing. *Science*, 335(6068):542–543, 2012.

## A. MisciteBench Construction Details

### A.1. Source Selection Rationale

Our goal in constructing MISCITEBENCH is to obtain a contamination-controlled, cross-disciplinary benchmark that (i) represents the *actual* scientific articles that structure their fields and (ii) admits rich downstream citation structure for the Evidence Committee in ICSV. This motivates our choice of “most-cited article in the JIF-leading journal per JCR subject category” rather than random sampling.

Concretely, for each of the 254 Clarivate JCR subject categories, we perform the following steps:

1. **Journal selection.** We identify the 2024 Journal Impact Factor (JIF) leader within the subject category. When multiple journals share identical JIF up to three decimal places, we break ties by (i) total citable items and then (ii) alphabetical order of journal title.
2. **Article-level selection.** Within the JIF-leading journal, we consider all “research article”-type items published in 2024–2025.<sup>1</sup> Among these, we select the most-cited paper according to citation counts as of 2025-11-30. Citations are computed within the same JCR subject category to avoid advantaging articles that cross over into much larger neighboring fields.
3. **Eligibility checks.** We discard candidate sources that (a) are not primarily original research reports (e.g., survey/overview/review-type articles, tutorials, conference overviews, and other narrative syntheses), (b) are direct translations or republications of older work, or (c) have fewer than 5 downstream citations recorded in curated citation indexes (required for ICSV to assemble a non-trivial Evidence Committee).

This procedure offers three advantages. First, it guarantees that each source paper is a *field-central* object around which substantial citation behavior has already accumulated, rather than a marginal or rarely cited work. Second, by choosing within each JCR category, we avoid overrepresenting large or fast-growing fields and maintain coverage across 21 high-level disciplines. Third, anchoring the time window in 2024–2025 maximizes exposure to contemporary citation practices and reduces the risk that frontier LLMs have memorized the full text, which is critical for the knowledge-blank protocol in Appendix A.2.

### A.2. Knowledge-Blank Cleanroom Fact-Checking Protocol

The knowledge-blank “cleanroom” protocol is designed to ensure that evaluated models cannot rely on parametric memory of source papers in MISCITEBENCH, but must instead reason over the documents they are given at evaluation time. Intuitively, a paper enters MISCITEBENCH only if a broad panel of strong models collectively “fail” to recognize its internal details, even when explicitly prompted to recall them.

**Fact-check probe synthesis via LRM + human curation.** For each candidate source paper, we construct  $N = 10$  forensic *fact-check questions* such that answering them correctly requires access to the main body or appendices and cannot be solved from title, authors, venue, year, or general background knowledge alone. We follow these design guidelines:

- Questions target *specific*, non-obvious details: numerical results, ablation outcomes, limitations, cross-subgroup comparisons, key hyperparameters, or subtle qualitative caveats.
- Each question has a deterministic, single ground-truth answer (a number, short phrase, or Boolean) so that correctness is well-defined.
- The answer is supported by a single gold snippet that a human auditor can locate in the full text in a short time.
- Questions are balanced in difficulty, mixing relatively simple pattern-completion probes (e.g., filling in a key number) with more intricate checks that hinge on fine-grained details.
- All questions must be grounded in the *content of the source paper itself*, not in any paper that the source cites.

<sup>1</sup>We exclude editorials, letters, corrigenda, and news-and-views pieces by filtering on publisher-provided article-type metadata and, when ambiguous, by manual inspection.

To construct these probes, we first apply a Large Reasoning Model (LRM; gemini-2.5-pro) to the *full text* of the candidate paper, including appendices. The LRM is instructed to read the paper sentence by sentence, perform deep reasoning to understand every detail, and then propose 15 candidate fact-check questions together with their canonical answers and supporting snippets. The exact prompt we use is:

### Fact-check question synthesis prompt

You will be given the full text of a scientific paper (including any appendices). Read the paper carefully, line by line, and use very deep reasoning to ensure that you understand every detail of its content.

Your task is to design 15 fact-check (FC) questions about this paper that can be used to test whether a large language model has memorized the paper as training data and truly knows its internal details.

#### Requirements:

1. Each FC question must have a single, deterministic ground-truth answer. The answer must not be vague, subjective, or open-ended. It should be short (e.g., a specific number, phrase, or Boolean) so that correctness can be checked reliably.
2. The FC questions must probe details that are specific to this paper. Avoid anything that could be guessed from common sense or generic background knowledge. Focus on paper-exclusive details such as exact numbers, specific experimental configurations, particular subsets, ablation outcomes, or precise verbal caveats.
3. The set of 15 FC questions should cover a balanced range of difficulty:
  - some easier checks (e.g., completing a key number in a sentence),
  - some medium-difficulty checks,
  - some harder checks that require integrating information across sections (e.g., main text + appendix).All questions must be strictly faithful to the paper; do not introduce hallucinated content.
4. Prefer high-memorization cues: if this paper had been used as part of a training set, a model that memorized it should be able to answer these questions accurately.
5. Very important: all FC questions must be derived from the content of this paper itself. Do NOT base any question on the content of other papers cited in the references.

#### Output format:

Return the 15 FC questions in standard CSV form with the following columns:

- "Paper Title": the exact title of this paper in English.
- "FC Question": the fact-check question in English.
- "Standard Answer": the unique correct answer in English.
- "Original Text": the exact snippet from the paper where the answer appears.

For any mathematical expressions in "Standard Answer" or "Original Text", use inline LaTeX with  $\$...$$  and ensure that the formulas are completely correct and compilable.

All content in the CSV must be in precise English.

#### Paper full text:

[INSERT PAPER HERE]

From the 15 LRM-generated candidates, human annotators then select and edit a final set of  $N = 10$  questions per

paper. During curation, annotators (i) verify that each question truly requires full-text access, (ii) remove any probe that could plausibly be answered from metadata or general domain knowledge, and (iii) correct the Standard Answer and Original Text fields when necessary.

**Gating models and prompt.** We probe a fixed, heterogeneous panel of frontier and strong open models: gpt-4o-2024-08-06, o4-mini-2025-04-06, gpt-5-2025-08-07, claude-sonnet-4-20250514, claude-opus-4-20250514, gpt-oss-120b, gpt-oss-20b, Nemotron 3 Nano (30B A3B), Llama-3.3 Nemotron Super 49B v1, Qwen3-235B-A22B-Thinking-2507, Qwen3 VL 32B Thinking, Qwen3 VL 8B Thinking, Minstral 3 (14B Reasoning 2512), Magistral Medium, Deepseek-V3.2 (Thinking), DeepSeek-R1-0528, Deepseek R1 Distill Qwen 32B, Deepseek R1 Distill Qwen 14B, Llama 3.1 405B Instruct, and Llama 3.3 70B Instruct.

For each model, we deliberately restrict access to *metadata only*: title, author list, journal/venue, and publication year. The abstract and full text are *never* provided to the gating models. We explicitly disable tool usage and web browsing.

Each model receives, for each question, the following fact-check gate prompt.

#### System Prompt for Fact-Check Gate

You are evaluating whether you already know the internal contents of a scientific article from your training data.

You will see ONLY the article's title, authors, journal/venue, and year.  
You will NOT see the abstract or full text.

Then you will be asked a highly specific factual question about the article.

Very important:

- Do NOT guess.
- If you are NOT certain of the exact answer based on what you already know about this paper from pre-training, answer with the single token: UNKNOWN.
- Only answer with a concrete value if you are sure it exactly matches what appears in the original paper.

#### User Prompt for Fact-Check Gate

Article metadata

Title: [TITLE]

Authors: [AUTHORS]

Journal: [VENUE]

Year: [YEAR]

Fact-check question

[QUESTION\_i]

Instructions

If you are certain you know the exact answer from your internal knowledge of this paper, answer with a short phrase or number.

Otherwise, answer with the single token: UNKNOWN.

**Automatic grading and acceptance criterion.** For each question, we store the canonical gold answer  $a_i$ . Let  $r_i^{(m)}$  denote model  $m$ 's response to question  $i$ .

- **Numeric answers.** If  $a_i$  is numeric, we parse  $r_i^{(m)}$  as a real number whenever possible and count it as correct if

$$\frac{|r_i^{(m)} - a_i|}{\max(1, |a_i|)} < 0.01.$$

- **Textual answers.** If  $a_i$  is textual, we do not rely on brittle string equality. Instead, we use an independent LLM grader (gpt-4o-2024-08-06) that receives both the gold answer and the model’s response and decides whether they are *semantically equivalent*. The grading prompt is:

#### Grading Prompt

You are grading whether two short answers express the same factual content.

Gold answer: "[GOLD\_ANSWER]"  
 Model answer: "[MODEL\_ANSWER]"

If, ignoring superficial differences in wording, the two answers refer to the same specific fact with the same level of specificity, respond with YES.  
 Otherwise respond with NO.

Respond with exactly one token: YES or NO.

A textual response  $r_i^{(m)}$  is marked correct if and only if the grader returns YES.

A candidate paper is *admitted* into MISCITEBENCH if and only if every model in the panel fails all probes, i.e.,

$$\forall m \in \mathcal{M}, \quad \sum_{i=1}^N \mathbf{1}[r_i^{(m)} \text{ is correct}] = 0.$$

If any model answers at least one probe correctly, we discard the paper and move to the next most-cited article in the same journal and time window, repeating the procedure until a clean paper is found or the candidate pool is exhausted.

### A.3. Dual-Tier Adversarial Miscitation Generation and Validation

This section details how we generate and validate adversarial miscitations grounded in the five-category taxonomy described in Section 2.1. The goal is to create instances that are (i) realistic, (ii) taxonomy-pure (each miscitation belongs unambiguously to one category), and (iii) calibrated in difficulty, ranging from surface-level errors to expert-level traps.

**LRM generation prompt and output schema.** Given a source paper  $B$ , we provide a Large Reasoning Model (gemini-2.5-pro) with the *full* paper, including all main sections and appendices, together with the detailed definitions of the five miscitation categories. The model is instructed to first perform extremely deep reading and then, for each category, synthesize both moderate and very difficult miscitations.

The core generation prompt (simplified for presentation) is:

#### Prompt to LRM for miscitation synthesis.

We are constructing a dataset of diverse miscitation examples to evaluate automatic miscitation detection systems.

You will be given the FULL TEXT of a source paper (including appendices) and the definitions of 5 miscitation categories. First, read the paper slowly and carefully, sentence by sentence, and use very deep reasoning to ensure that you understand its methods, results, limitations, and nuanced details.

Then, for EACH of the 5 miscitation categories, you must create 5 miscitation

examples that (incorrectly) cite this paper:

- 3 examples should be single sentences or short paragraphs that are clearly wrong for a careful reader who deeply understands the paper, but not so trivial that they can be spotted at a glance.
- 2 examples should be long, complex sentences or slightly longer paragraphs that are VERY difficult to detect as wrong. These "deep" miscitations should rely on subtle, easily overlooked details of the paper, such that even domain experts or the original authors would need to think carefully before identifying the error.

Additional constraints:

- Different scenarios under the same category should be diverse. Cover as many distinct ways of committing that type of miscitation as possible.
- Different categories must remain well separated. Each designed example should clearly belong to exactly ONE miscitation category under the taxonomy, with no ambiguity or overlap.
- For EVERY miscitation example, you must also provide:
  - \* Explanation: a clear English explanation of why this is a miscitation and how it instantiates the target category.
  - \* Correct Statement: a corrected citing sentence or short passage that would describe the source paper accurately.
  - \* Original Text: the exact snippet(s) from the source paper that grounds your judgment (i.e., what the paper actually says).
- Read the source paper deeply enough to avoid being misled by possible typesetting or formatting errors.
- Very important: miscitation content must be about THIS source paper. Do NOT base any miscitation on the content of other papers cited in its references.

Output format:

Return all 25 designed miscitations (5 categories \* 5 examples) in CSV form with the following columns:

- "Miscitation": the incorrect citing sentence or short paragraph.
- "Explanation": why this is a miscitation, in clear English.
- "Correct Statement": a corrected version that would cite the paper properly.
- "Original Text": the supporting snippet(s) from the source paper.
- "Miscite Type": one of {Citation Validity Error,  
Content Misrepresentation Error,  
Scope Extrapolation Error,  
Evidence Characterization Error,  
Attribution & Traceability Error}.
- "Difficulties": one of {SURFACE, DEEP}, where SURFACE indicates a moderate, more local error, and DEEP indicates a very difficult, globally subtle error.

All content must be in precise English. Ensure that each example strictly matches its assigned miscitation type and difficulty level.

Source paper full text:

[INSERT PAPER HERE]

Miscitation taxonomy:

[INSERT 5 CATEGORY DEFINITIONS HERE]

This prompt yields, for each source paper, 25 candidate miscitation rows covering the full taxonomy: three SURFACE

and two DEEP instances per category, with explicit fields `Miscitation`, `Explanation`, `Correct Statement`, `Original Text`, `Miscite Type`, and `Difficulties`. We perform schema validation and reject any row where the generated `Miscite Type` is inconsistent with the textual description in `Explanation`.

**Post-processing and CSV representation.** The raw LRM outputs are parsed into a standardized CSV representation. We normalize category labels to the five taxonomy names, collapse minor variations in difficulty tags into SURFACE vs. DEEP, and enforce basic well-formedness constraints (non-empty fields, length limits, and valid category labels). Instances that appear to mix multiple error types (e.g., both misattribution and content distortion) are flagged for later removal or manual repair.

**Independent LRM consistency check.** Each candidate miscitation is then re-evaluated by an *independent* LRM with a different architecture and training history (`gpt-5.1-Extended-Thinking`). For each row, the independent model receives:

- the full source paper,
- the candidate `Miscitation` sentence/paragraph,
- the proposed `Miscite Type` and `Difficulties`,
- the `Original Text` snippet, and
- the original `Explanation` and `Correct Statement`.

The independent LRM is asked to (i) decide whether the candidate is in fact a miscitation of the source, (ii) assign the most appropriate taxonomy category, and (iii) provide its own free-form explanation of the error and a corrected citing statement.

We then apply two automatic filters:

- **Category agreement.** If the independent model assigns a different primary miscitation type than the original `Miscite Type`, the instance is flagged for removal or manual repair.
- **Explanation alignment.** We again use `gpt-4o-2024-08-06` as a semantic grader to compare the independent model’s explanation to the original `Explanation`. Only instances where the two explanations are judged to be semantically equivalent (same mechanism, same primary error) are retained.

Examples that fail either check are either edited and re-validated or dropped entirely.

**Human expert validation.** Finally, we subject every source paper and its 25 candidate miscitations to rigorous human review by domain experts. We recruit a panel of PhD students, researchers, and senior practitioners spanning all 21 high-level disciplines covered by the 254 JCR subject categories. Collectively, their expertise covers the full set of subfields represented in `MISCITEBENCH`.

For each source paper, we assign 3–4 experts whose research area matches the corresponding high-level discipline (e.g., Clinical Medicine, Computer Science, Social Sciences). Each expert is given the full source paper and the 25 miscitation rows and asked to perform cross-validation under the following criteria:

1. **Is it truly a miscitation?** Does the `Miscitation` sentence or paragraph in fact misrepresent the source paper, and does the stated `Explanation` accurately and fully capture the reason it is wrong?
2. **Taxonomy alignment.** Given the definitions in our five-category taxonomy, does the assigned `Miscite Type` correspond to the *primary* error mechanism? Could a reasonable annotator confidently place this instance in exactly this category?
3. **Evidence grounding.** Does the `Original Text` field correctly and sufficiently point to the span(s) in the source paper that justify the error diagnosis (and, for corrected statements, their validity)?

4. **Corrected statement accuracy.** Is the `Correct Statement` factually consistent with the source paper and free from new misrepresentation issues?

A misrepresentation instance is retained in MISCITEBENCH only if all assigned experts agree that it satisfies the criteria above. If any expert disagrees on misrepresentation status, taxonomy label, evidence grounding, or corrected-statement accuracy, the instance is either revised and re-submitted to the same validation protocol or discarded.

After this multi-stage pipeline—LRM generation with deep reading, independent cross-model consistency checking, and stringent domain-expert cross-validation—we obtain the final set of 254 source papers  $\times$  5 categories  $\times$  5 instances per category, for a total of 6,350 high-quality misrepresentation cases used in MISCITEBENCH.

## B. Taxonomy Annotation Protocol

This appendix provides the full details of the expert annotation study referenced in Section 2.1. The goals of this study are twofold: (i) to test whether the five-category taxonomy is empirically usable and complete across disciplines, and (ii) to quantify how reliably independent experts can reproduce each other’s judgements when constrained by the Dependency Precedence Rule.

### B.1. Sampling Strategy and Annotator Panel

**Instance selection.** From the full MISCITEBENCH benchmark of 6,350 misrepresentation instances, we draw a stratified sample of  $N = 500$  cases for manual validation. Stratification is performed jointly over:

1. **Taxonomy category:** we maintain approximately equal representation of the five error types (Citation Validity, Content Misrepresentation, Scope Extrapolation, Evidence Characterization, Attribution & Traceability).
2. **Discipline:** we preserve the distribution over the 21 high-level disciplines used in MISCITEBENCH construction (e.g., Agricultural Sciences, Clinical Medicine, Computer Science, Social Sciences, Visual & Performing Arts), so that no single field dominates the sample.

This design ensures that the annotation study probes both the breadth of scientific domains and the full error-code space, rather than concentrating on a narrow subset of easy or homogeneous instances.

**Annotator qualifications.** Each instance is independently labeled by 3 experts with doctoral-level or advanced graduate training in a subfield relevant to the source paper. We restrict assignment so that:

- every annotator has prior experience with reading and evaluating peer-reviewed scientific articles in the corresponding discipline, and
- no annotator is asked to label instances from a field outside their demonstrated area of expertise.

Experts are blind to the synthetic origin of MISCITEBENCH and are instructed to treat each instance exactly as they would when reviewing a real manuscript.

### B.2. Annotation Materials and User Interface

For each candidate misrepresentation instance, annotators are provided with a compact but fully contextualized bundle:

- **Source-side context:** the title, abstract, and gold supporting excerpt (i.e., the span in the source paper that the misrepresentation was constructed to distort).
- **Citing-side context:** the expanded citing window  $S_{\text{expanded}}$  (the citing sentence plus its immediate neighbors), matching the context used by BIBAGENT during verification.
- **Taxonomy reference:** the definitions of the five categories, including their diagnostic litmus questions and short examples, identical to those in Section 2.1.

The annotation interface is a structured form that guides experts through the same logical decision path used by BIBAGENT. For each instance, the interface:

1. presents the citing context and source excerpt side-by-side,
2. displays the five taxonomy categories with their litmus questions in a collapsible panel, and
3. enforces the Dependency Precedence Rule via a hierarchical selection widget (described below).

### B.3. Decision Procedure and Dependency Precedence

Annotators are explicitly instructed to mimic the stepwise verification process a careful reviewer would perform. The interface enforces the following sequence:

1. **Attribution & Traceability check.** Determine whether, given the citation metadata, a reader could reliably locate and identify the correct source (Attribution & Traceability Error). If an Attribution failure is present (e.g., ghost citation, hopelessly ambiguous metadata), annotators select this label and no further categories can be chosen.
2. **Citation Validity check.** If the citation is traceable, decide whether the source itself remains valid as scientific evidence (Citation Validity Error), e.g., retracted or misused secondary source. If so, this label is selected and lower levels are disabled.
3. **Content Misrepresentation check.** If the source is valid, compare the citing context to the source excerpt and decide whether the citing text faithfully represents the factual content (Content Misrepresentation Error).
4. **Scope Extrapolation check.** If content is correctly represented, decide whether the citing text applies an otherwise valid conclusion outside the population, setting, or task for which it was established (Scope Extrapolation Error).
5. **Evidence Characterization check.** Finally, decide whether the citing text mischaracterizes the logical type or strength of the evidence (Evidence Characterization Error), e.g., treating correlational results as causal or overstating statistical certainty.

This procedure operationalizes the *Dependency Precedence Rule* described in Section 2.1: once an annotator records a failure at a higher level (e.g., Attribution), all lower levels (Validity, Content, Scope, Evidence) are automatically grayed out and cannot be selected for that instance. Conversely, if no failure is detected at any level, the annotator records the instance as *No miscitation*.

In addition to the five primary taxonomy labels, the interface provides two auxiliary options:

- **Other:** the annotator judges that there is a genuine miscitation, but it does not fit any of the five categories.
- **Uncertain:** the annotator cannot confidently decide due to ambiguity or insufficient information.

For each chosen label (including “Other” and “Uncertain”), annotators supply a short free-text rationale explaining their decision. These rationales are later used during adjudication and for qualitative error analysis.

### B.4. Agreement Measurement

To quantify reproducibility, we compute pairwise Cohen’s  $\kappa$  among the three annotators over the five taxonomy labels only, ignoring instances labeled as “Other” or “Uncertain” by any annotator.<sup>2</sup> For annotators  $a$  and  $b$ , Cohen’s  $\kappa$  is defined as:

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \quad (14)$$

where  $p_o$  is the observed label agreement and  $p_e$  is the expected agreement under chance, computed from the empirical label marginals.

---

<sup>2</sup>We exclude “Other” and “Uncertain” from the primary agreement analysis because they are by design fallback categories; their frequency is reported separately.

Across annotator pairs, we obtain an average  $\kappa = 0.73$ , which falls in the “substantial” agreement range under conventional benchmarks. This indicates that, given the same source and citing contexts and a shared taxonomy, independent experts largely converge on the same primary error category.

The marginal usage rates of the fallback options are low: “Other” is selected in 4.2% of annotations and “Uncertain” in 3.6%. These small rates, together with the substantial  $\kappa$ , provide empirical evidence that:

1. the five-category taxonomy is sufficiently expressive for the MISCITEBENCH corpus, and
2. the operational litmus questions make category boundaries clear enough for reproducible annotation.

### B.5. Adjudication and Final Labels

After individual annotation, we perform a structured adjudication step to obtain a single gold label per instance:

- **Simple majority.** If at least two annotators agree on the same taxonomy category, that category is taken as the final label.
- **Complete disagreement or fallback use.** If all three annotators select different categories, or if any annotator chooses “Other” or “Uncertain”, the instance is escalated to a fourth senior annotator.

The senior annotator has full access to:

1. the original annotation bundle (source excerpt, citing context, taxonomy definitions),
2. the three annotators’ labels, and
3. their free-text rationales.

They then assign a final label consistent with the Dependency Precedence Rule, optionally revising the instance text if a clear typographical or formatting error is discovered. In practice, adjudication rarely required reconsidering the underlying taxonomy itself: no new error categories were requested, and all disputed cases could be resolved by clarifying which litmus question applied most directly.

Taken together, the low reliance on “Other” and “Uncertain”, the substantial inter-annotator agreement, and the lack of new categories during adjudication support our central claim: the proposed five-category taxonomy is both *complete* for the failure modes captured in MISCITEBENCH and *operationally precise* enough to be applied reproducibly across disciplines.

## C. Document Parsing and Citation Mapping Details

The Document Parser and Citation Mapper (DPCM) normalizes heterogeneous inputs (L<sup>A</sup>T<sub>E</sub>X, XML/HTML, and PDFs) into a single hierarchical Markdown representation. This representation preserves the discourse skeleton of the paper—headings, paragraphs, math, figures/tables, and citation anchors—so that downstream verification always operates on a traceable, format-agnostic substrate. This appendix details (i) markup-based normalization, (ii) PDF parsing via image-based layout serialization, (iii) the Extraction Verifier, and (iv) citation mapping.

### C.1. Markup-Based Normalization

For L<sup>A</sup>T<sub>E</sub>X and XML/HTML inputs, DPCM constructs an intermediate abstract syntax tree (AST) and then projects it into Markdown while deliberately retaining only semantics that matter for citation auditing.

**Macro handling and structural projection.** We expand standard inline macros such as `\emph`, `\textbf`, and `\textit` where they affect emphasis, and prune purely presentational directives (spacing, color, line breaks) that do not change logical structure. Sectioning commands (`\section`, `\subsection`, etc.) are converted into Markdown headings (#, ##, ##), while lists, display math, and inline math are serialized as Markdown lists and  $\dots$  blocks. Cross-references (`\ref`, `\eqref`, etc.) are preserved as literal text so that equation, figure, and table numbering can later be checked for consistency.

**Citation rendering and anchoring.** Citation commands in markup (`\cite`, `\citet`, `\citet` and style-specific variants) are processed in two coupled layers:

1. *Surface rendering.* Using the document’s bibliographic style configuration (e.g., numeric, author–year/APA, Chicago), we re-render each citation into the same surface form that a compiled PDF would produce: for example, `\citet{Smith2020,Brown2019}` becomes “(Smith, 2020; Brown & Brown, 2019)” under APA-style settings or “[12, 19]” under a numeric style. As a result, the Markdown closely mirrors the citing text as seen by human readers, which is crucial for later LLM-based reasoning.
2. *Stable internal anchors.* In parallel, DPCM attaches an internal anchor to each rendered citation that records the originating citation keys and their order. These anchors are *not* shown in the surface Markdown, but they form a lossless mapping from each inline citation occurrence back to the corresponding bibliography entries. Citation mapping (Section C.4) later consumes these anchors to build sentence-level citation graphs, while ACSV and ICSV operate purely on the human-readable surface text.

The result of markup normalization is a hierarchical Markdown document that faithfully reflects the compiled reading experience, while still exposing a machine-tractable citation backbone.

## C.2. PDF Parsing via Image-Based Layout Serialization

Text-based PDF extraction often yields broken reading order, missing math, and unreliable treatment of figures and tables. For PDFs, DPCM therefore uses an explicit image-based pipeline that delegates layout reconstruction to a multimodal model and then repairs cross-page boundaries.

### C.2.1. PAGE RENDERING AND PAGE-LEVEL TRANSCRIPTION

Each PDF is first converted into one raster image per page at a fixed resolution of 300 DPI using `pdf2image`. The parser does not rely on embedded PDF text; instead, each page image is passed directly to `gpt-4o-2024-08-06` together with a concise system prompt that specifies the Markdown serialization rules.

The system prompt used for page-level transcription is:

```
Please accurately convert the main body text in the image into markdown.
Use $...$ for inline formulas and $$...$$ for display formulas.
```

Requirements:

1. Preserve the original wording exactly. Do NOT paraphrase, add, or omit text.
2. Extract only main body text.
  - Ignore illustrations and graphical elements.
  - Keep captions whose text starts with "Figure", "Table", "Fig.", or "Tab".
3. If the image contains tables, convert them faithfully into markdown tables.
4. Remove headers, footers, page numbers, and sidebars that are not part of the main body.
5. Do not call any external tools; rely only on your visual understanding.
6. If the image contains no readable text, output: null

The model returns Markdown for the page; if it wraps the content inside a code fence, we strip the fence and retain only the inner Markdown. This gives us a sequence of per-page Markdown fragments that already approximate the correct reading order, with headings, paragraphs, displayed math, inline math, figure and table captions, and inline citations rendered as they appear on the page.

### C.2.2. CROSS-PAGE BOUNDARY REPAIR

Page-level OCR and transcription inevitably break some sentences across page boundaries. DPCM therefore applies a dedicated cross-page repair pass before any structural verification or citation mapping.

We maintain two parallel views of each page: (i) the original page-level Markdown output, and (ii) a tagged version used for

debugging, in which we mark potentially incomplete beginnings or endings using lightweight heuristics.

**Incomplete head/tail tagging.** For each page, we inspect the first and last non-empty lines. A line is considered a *complete* tail if it ends with sentence-final punctuation (e.g., “.”, “!”, “?” or their CJK counterparts), closes a math block (\$ or \$\$), ends with a citation or reference pattern (e.g., “[12]”), or is a plausible standalone heading or caption. We consider the first content line to be a *complete* start if it looks like the beginning of a new sentence or block (e.g., capitalized heading, list item, or math environment) rather than a continuation.

If the tail of a page fails these checks, we tag it with <INCOMPLETE-END\_Pn>; analogously, if the head of the next page looks like a continuation (lowercase start, conjunctions such as “and”, “but”, “however”, leading bracket, etc.), we tag it with <INCOMPLETE-START\_Pn>. These tags do not change the content but mark candidate boundaries for repair.

**Heuristic merging and hyphen repair.** We then merge per-page texts into a single document using a boundary-aware procedure. For each adjacent page pair, we:

- Split each page into paragraphs separated by blank lines.
- Separate figure/table captions from main text, so that caption-only paragraphs are appended without being merged into running prose.
- Examine the last main-text paragraph of page  $n$  and the first main-text paragraph of page  $n+1$ .

If the tail paragraph is incomplete and the head paragraph does not resemble a new heading or section, we attempt to repair the boundary:

1. *Hyphenated words.* If the tail ends with a hyphenated fragment (e.g., “multi-” at the end of a line) and the head begins with alphabetic characters, we merge them into a single word (e.g., “multi-agent”), keeping a single hyphen when appropriate.
2. *Direct continuation.* If the head starts with a lowercase letter, discourse connective (*and, but, however, therefore, which, that, because*, etc.), or an opening bracket, we join the two paragraphs with a space, treating them as a single sentence that crossed the page break.

If neither heuristic confidently applies but the tail paragraph is clearly incomplete and not a caption, we fall back to a minimal LLM-based boundary repair step that edits only the ambiguous junction.

**LLM-based boundary repair.** For hard cases, we send the last paragraph of page  $n$  and the first paragraph of page  $n+1$  to gpt-4o-2024-08-06 with a dedicated system prompt:

You repair cross-page sentence breaks in markdown text extracted from a scientific PDF. You are given:

- PREV: the tail paragraph of page N
- NEXT: the head paragraph of page N+1

They may contain:

- a sentence split across the page break,
- duplicated fragments, or
- a word split by a hyphen.

Your task:

- Minimally fix the boundary so that the text reads as a single continuous document.
- Do NOT invent new content or change meaning.
- Do NOT drop valid content except for exact duplicates.
- You may:

- move tokens between PREV and NEXT,
- remove duplicated fragments,
- join hyphenated words split across the boundary.

Output exactly in this format:

```
PREV_FIXED:  
<fixed text for the previous-page tail>  
---  
NEXT_FIXED:  
<fixed text for the next-page head>
```

We parse PREV\_FIXED and NEXT\_FIXED from the model output and adopt them only if the boundary has changed in a way that improves continuity (e.g., duplication removed, sentence completed). Otherwise, we keep the original paragraphs. This design ensures that the LLM operates as a local repair operator, not as a free-form rewriter of the document.

After iterating over all boundaries, we remove debug tags <INCOMPLETE\_START\_Pn> and <INCOMPLETE\_END\_Pn> and normalize whitespace, yielding a single merged Markdown file that reflects the visual layout and reading order of the PDF while repairing cross-page sentence breaks.

### C.3. Extraction Verifier

Once the merged Markdown is available, the Extraction Verifier performs a multi-level audit to detect structural and bibliographic inconsistencies before citation reasoning. The verifier combines deterministic checks with targeted LLM-based inspection.

**Level 1: deterministic structural checks.** We parse the Markdown into a lightweight document tree and enforce basic invariants:

- *Heading progression.* Heading levels must not jump by more than one level (e.g., # → ## → ### is allowed; # → ### is flagged). Abrupt regressions (e.g., ### after #) are similarly flagged.
- *Equation and float numbering.* Equation labels (“(1)”, “(2)”, …) and figure/table numbers (“Figure 3”, “Table 2”) are checked for monotone, mostly contiguous sequences. Small gaps are tolerated, but long stretches of missing or duplicated numbers trigger a warning.
- *Citation index sequences.* For numeric styles, we extract all integers appearing in citation-like patterns (e.g., “[3]”, “[3, 5, 9]”, “[3–5]”) and verify that the global sequence is roughly monotone and dense. Large gaps (e.g., no references between [5] and [20]) or frequent out-of-order jumps signal potential parsing failures, such as skipped bibliography segments or mis-ordered text blocks.

For each violation, the verifier records the surrounding lines and the page range where the anomaly occurs.

**Level 2: localized re-parsing.** When Level 1 identifies a likely structural discontinuity that aligns with a small set of pages (e.g., a break in equation numbers between pages 7 and 8), we re-run the PDF-to-Markdown conversion only on those pages. The re-parse uses the same 300 DPI images and system prompt as in Section C.2, but the new output is compared against the original to detect missing lines, duplicated blocks, or altered reading order. If the re-parsed fragment resolves the anomaly (e.g., restores a missing equation or bibliography block), we splice it into the global document; otherwise, we keep the original and proceed to Level 3.

**Level 3: LLM-based semantic audit.** For residual suspicious regions where deterministic rules cannot confidently decide (e.g., text density drops abruptly without clear structural markers), we invoke a light semantic audit. We provide gpt-4o-2024-08-06 with the Markdown snippet around the anomaly and a succinct description of the expected structure, asking whether the snippet appears truncated, duplicated, or out-of-order. A representative prompt is:

You are auditing markdown extracted from a scientific paper.

You are given a short markdown segment and a brief description of the surrounding document structure (e.g., "between Section 3.2 and Section 3.3").

Decide whether the segment shows signs of extraction error:

- missing lines or sentences,
- duplicated paragraphs,
- obvious reading-order errors (e.g., a caption in the middle of a sentence).

Answer with one of:

- OK
- SUSPICIOUS\_MISSING
- SUSPICIOUS\_DUPLICATE
- SUSPICIOUS\_ORDER

Then, in one short sentence, explain your choice using only evidence that is visible in the provided markdown.

Segments labeled as suspicious are either re-parsed again with tighter crops or flagged as low-confidence regions for downstream modules (ACSV/ICSV) and, if needed, for manual inspection. In practice, this three-level verifier reduces gross extraction failures (e.g., missing sections, mis-ordered columns) while keeping the pipeline predominantly deterministic and auditable.

#### C.4. Citation Mapping

Given a structurally verified Markdown document, the citation mapping component constructs a sentence-level citation graph that is consumed by CSAC, ACSV, and ICSV. The goal is to map each inline citation span in the body text to one or more normalized bibliography entries, while preserving the original surface style.

**Inline citation detection.** We first identify citation spans in the Markdown using style-agnostic patterns. The detector considers three families:

- *Numeric citations*, such as “[12]”, “[3, 5, 9]”, “[3–5]”, and inline variants like “(see also [7])”.
- *Author–year citations*, such as “(Smith, 2020)”, “Smith (2020)”, and grouped forms like “(Smith, 2020; Brown & Lee, 2019)”.
- *Footnote-based references*, where the citation marker appears as a superscript or bracketed index in the main text and the full reference appears in a footnote block.

By scanning the entire document, we infer the dominant citation style (numeric vs. author–year vs. note-based) from frequency statistics and pattern coverage, and adapt parsing rules accordingly. Each detected span is associated with its containing sentence (using punctuation-based segmentation with list/heading safeguards), yielding a preliminary mapping from sentences to unnormalized citation strings.

**Bibliography parsing.** We then isolate the bibliography section (or sections) by detecting headings such as “References”, “Bibliography”, or style-specific variants. Each bibliographic entry is parsed into a normalized record containing at minimum: canonical author list (last names and initials), publication year, title tokens, venue/journal name, and any explicit identifiers (DOI, arXiv ID, PubMed ID). For markup-based inputs, we prefer the original `\bibitem` or BibTeX metadata; for PDF-only inputs, we rely on typography cues (hanging indentation, numbering, bullet markers) combined with pattern-based parsing.

**Citation-to-entry alignment.** Finally, we align inline citation spans to bibliography entries:

- *Markup inputs.* When the document originates from L<sup>A</sup>T<sub>E</sub>X or XML/HTML, we use the internal anchors described earlier to directly map each inline citation occurrence to a set of bibliography records, with no string matching required. This mapping is exact and order-preserving.
- *PDF-only inputs, numeric style.* We map numeric indices in citations (e.g., “[7]” or “[3–5]”) to positions in the parsed bibliography list, taking into account style-specific conventions (e.g., whether numbering restarts in supplements). Ranges and grouped citations are expanded into individual edges (e.g., “[3–5]” yields links to 3, 4, and 5).
- *PDF-only inputs, author–year style.* For each citation string, we extract last names and year(s), then compute a similarity score between this tuple and each candidate bibliography entry, based on overlap of normalized last names, year equality, and title token similarity. The highest-scoring candidate above a threshold is selected; ties and sub-threshold cases are flagged as ambiguous.

In ambiguous or noisy cases (e.g., partial author lists, missing years), we use a lightweight LLM-assisted disambiguation step: we provide the surface citation string and a small set of candidate bibliography entries and ask the model to select the best match or to abstain if none is appropriate. This step is constrained to choose from explicit candidates and does not fabricate new references.

The final output of citation mapping is, for each sentence including intext citation(s) in the document, a set of resolved citation edges pointing to normalized metadata records (title, authors, year, venue, DOI). This structure is the entry point for CSAC, which determines source accessibility, and for ACSV/ICSV, which perform taxonomy-aligned mispuration verification on top of a fully traceable citation graph.

## D. ICSV Implementation Details

This appendix documents the concrete implementation of the **Inaccessible Cited Source Verifier (ICSV)** and its **Evidence Committee** mechanism. ICSV targets the strict paywall regime: the full text of the cited source  $B$  is unavailable, so verification must be grounded in *auditable downstream evidence* rather than speculative reconstruction. We therefore (i) extract a *claim-preserving paraphrase* of what the citing paper  $A$  attributes to  $B$ , (ii) extract analogous attributions to  $B$  from multiple open-access downstream citers, (iii) organize these attributions into coherent aspects via LLM-based semantic clustering, (iv) assign field-normalized influence weights across heterogeneous venue types (journal / conference / preprint), and (v) compute a reliability-aware consensus verdict with calibrated confidence and principled abstention.

Unless otherwise noted, all ICSV LLM calls use `gpt-4o-2024-08-06`. We use `gpt-4o-2025-08-06` only for the semantic clustering step (Section D.4) to improve partition stability on long claim lists.

### D.1. Downstream Committee Retrieval and Witness Verification

Given a paywalled cited source  $B$ , CSAC provides a resolved metadata snapshot (DOI when available; otherwise title, author list, venue, year). ICSV then constructs an *open-access witness set*  $C_{\text{open}}$  of downstream citers that reference  $B$  and have retrievable full text.

**Primary citation-graph retrieval.** We query a curated open citation index (OpenAlex as primary; Crossref as DOI/metadata fallback) to enumerate works that cite  $B$ . We de-duplicate by DOI and canonical title normalization (lowercasing, punctuation stripping, whitespace collapse) and discard non-scholarly records (e.g., editorial notes) using venue/type metadata when available.

**Open-access eligibility and full-text acquisition.** A candidate citer  $p$  enters  $C_{\text{open}}$  only if (i) a full-text URL is available (publisher OA, institutional repository, or vetted preprint server), and (ii) the downloaded full text can be parsed by DPCM into structured Markdown with intact citation anchors.

**Witness validity check (must explicitly cite  $B$  in-text).** For each candidate citer  $p$ , we verify that  $p$  contains at least one *explicit in-text citation mention* to  $B$ . We accept a mention if any of the following match robustly: (i) DOI match, (ii) high-similarity title string match, or (iii) bibliography-entry match followed by an in-text anchor pointing to that entry. Candidates that do not pass this in-text witness check are discarded to prevent “false witnesses” created by noisy citation graphs.

**No premature skipping.** ICSV never “skips” a paywalled citation after retrieval. If witnesses are weak or contradictory, the system returns UNDECIDABLE with an explicit reliability diagnosis (Section D.7).

## D.2. Context-Aware Citing Claim Paraphrase (from Paper A)

The first step is to construct  $c_A$ : a *claim-preserving paraphrase* of what  $A$  attributes to  $B$  at the citation site. Importantly, our implementation does **not** ask the model to “rewrite the single main claim” (which invites summarization or abstraction). Instead, it performs a **tight paraphrase** of the cited sentence, with one permitted transformation: *resolve pronouns and implicit references into explicit mentions* while preserving all qualifiers, modality, and scope.

**Window expansion (iterative; no skipping).** Let  $s_A$  be the in-text sentence in  $A$  that cites  $B$ . We start from a minimal local window and expand only when necessary:

$$W_A(r) = \text{sent}(A, i - r) \oplus \cdots \oplus \text{sent}(A, i) \oplus \cdots \oplus \text{sent}(A, i + r),$$

where  $i$  is the index of  $s_A$  and  $r$  is the radius (default start  $r=1$ ). If the model returns INSUFFICIENT\_CONTEXT, we increment  $r \leftarrow r + 1$  and retry, continuing until a stable paraphrase is produced. We cap expansion by paragraph boundaries; if the paragraph is still insufficient, we extend to adjacent paragraphs in the same section.

**Stability criterion.** To avoid oscillating paraphrases under larger windows, we require *two consecutive radii* to produce paraphrases that are identical after normalization (whitespace collapse; punctuation normalization) or have semantic similarity above a high threshold (measured by a sentence embedding cosine similarity). The earlier paraphrase is then fixed as  $c_A$ .

**Prompt template (claim-preserving paraphrase).**

### System Prompt

You are an expert scientific copy-editor and verifier.  
 Your job is to produce a claim-preserving paraphrase of a specific sentence that cites  
     a prior paper B. Do NOT summarize, generalize, or add new claims.  
 Preserve all qualifiers (e.g., "may", "suggest", "in our setting"), all scope  
     constraints (population/task/conditions), and all numerical content.  
 The only required transformation is to resolve pronouns and implicit mentions into  
     explicit referents, using the provided context.

### User Prompt

```
[Context window W_A]
{W_A}

[Target sentence s_A that contains the in-text citation to paper B]
{s_A}

Instructions:
1) Paraphrase s_A as closely as possible (claim-preserving).
2) Resolve pronouns/implicit references using W_A (e.g., "this method" -> the named
   method).
3) Keep modality, qualifiers, and scope exactly faithful to s_A.
4) Do NOT mention citation markers, citation numbers, authors, or years.
5) Output ONE sentence only.
6) If the referent needed to resolve pronouns is not recoverable from W_A, output:
INSUFFICIENT_CONTEXT
```

### D.3. Witness Claim Extraction (from Each Downstream Citer)

ICSV extracts what each open-access witness paper  $p \in C_{\text{open}}$  claims about  $B$  at its explicit in-text citation sites.

**Selecting explicit mention sites.** For each witness  $p$ , we locate every sentence  $s_p$  whose in-text citation anchor resolves to  $B$  (via DOI/title/bibliography match as in Section D.1). For each such  $s_p$ , we run the *same* claim-preserving paraphrase procedure as Section D.2, producing a witness claim  $c_p^{(t)}$  for mention  $t$ . Within a witness paper, we de-duplicate claims using high-similarity filtering to avoid overweighting repeated boilerplate mentions.

**Outcome.** This yields a multi-set of witness claims

$$\mathcal{C}_B = \{(c_1, \text{src}(c_1)), \dots, (c_m, \text{src}(c_m))\},$$

where each  $c_\ell$  is a claim-preserving paraphrase attributed to  $B$  and  $\text{src}(c_\ell)$  denotes the source witness paper that produced it.

### D.4. LLM-Based Semantic Clustering and Evidence Statement Distillation

The goal is to organize  $\mathcal{C}_B$  into coherent “aspects” of  $B$  (e.g., method, dataset, empirical finding), then distill each aspect into a canonical evidence statement  $e_j$ . In our implementation, we **do not** use embedding-based agglomerative clustering, because choosing thresholds and cluster counts is brittle across fields and can merge distinct aspects of  $B$  into the same cluster. Instead, we perform **direct semantic clustering with an LLM**, which is both more controllable (via explicit constraints) and more robust to domain shift.

#### D.4.1. SEMANTIC CLUSTERING PROMPT (LLM CLASSIFIER)

We use gpt-4o-2025-08-06 to cluster witness claims into non-overlapping groups  $\{G_j\}_{j=1}^k$ . The prompt is designed to (i) prevent “semantic over-merging,” (ii) keep clusters interpretable, and (iii) enforce strict JSON output for deterministic parsing.

#### System prompt (semantic clustering).

You are an expert scientific librarian.  
You will receive a list of short, claim-preserving sentences that different papers attribute to a paywalled paper B.

Task: cluster these sentences into semantically coherent aspects of B.  
Each cluster must represent ONE aspect (e.g., one method contribution, one dataset, one key empirical finding, one theoretical claim).  
Do NOT merge two distinct aspects just because they are topically related.

#### Constraints:

- Every claim must belong to exactly one cluster.
  - Clusters should be as few as possible, but no cluster may contain claims that are substantively about different contributions/aspects.
  - If a claim is vague, place it into the closest cluster only if consistent; otherwise create a small "misc/unclear aspect" cluster.
- Return JSON only (no prose).

#### User prompt.

Paper B (metadata):  
Title: {title\_B}  
Year: {year\_B}  
Venue: {venue\_B}

Witness claims about B (each has an ID):  
1) {c\_1}

```

2) {c_2}
...
m) {c_m}

Output JSON with the following schema:
{
  "clusters": [
    {
      "cluster_id": "C1",
      "cluster_name": "short aspect name",
      "aspect_summary": "one-sentence description of the shared aspect",
      "claim_ids": [1, 7, 12]
    },
    ...
  ]
}

Rules:
- cluster_name must be <= 8 words.
- aspect_summary must be exactly one sentence.
- Do not invent content not present in the claims.

```

#### D.4.2. EVIDENCE STATEMENT DISTILLATION (PER CLUSTER)

For each cluster  $G_j$ , we distill a canonical evidence statement  $e_j$  that represents the shared content *as attributed by the community*. This step uses gpt-4o-2024-08-06.

##### System prompt (evidence distillation).

You are an expert scientific verifier.  
 You will see multiple claims that different papers attribute to a paywalled paper B about the SAME aspect. Your job is to write ONE canonical evidence statement that captures only their overlap.

##### Requirements:

- One sentence only.
  - Include critical qualifiers (scope, conditions, uncertainty).
  - Preserve numerical quantities if present and consistent.
  - If claims conflict, produce a conservative statement that reflects only what is common, and explicitly hedge (e.g., "is reported to", "suggests").
- Do NOT add new facts.

##### User Prompt.

Paper B (metadata):  
 Title: {title\_B}

Cluster {cluster\_id}: {cluster\_name}  
 Claims in this cluster:  
 - {c\_a}  
 - {c\_b}  
 ...

Write ONE sentence as the canonical evidence statement e\_j.

**Provenance.** For traceability, we store for each  $e_j$  the full set of contributing claim IDs and their source papers. All downstream weighting and voting operates on *unique witness papers* per cluster (Section D.5), preventing repeated mentions from the same paper from inflating support.

## D.5. Field-Normalized Influence Weights Across Journals, Conferences, and Preprints

Each evidence statement is only as credible as its witnesses. However, raw citations and venue prestige vary drastically across fields and publication years, and venue types differ (journals vs. conferences vs. preprints). We therefore compute a unified influence score  $\mathcal{I}(p)$  for each witness paper  $p$ , combining (i) **paper-level influence** within its field-year and (ii) **venue-level standing** within its field.

### D.5.1. PAPER-LEVEL CITATION PERCENTILE (ALL VENUE TYPES)

Let  $Cite(p)$  be the paper-level citation count from a citation index snapshot. We compute a field-year normalized percentile:

$$C_{\text{norm}}(p) = \text{Rank}_{\%}(Cite(p) \mid \text{Field}(p), \text{Year}(p)).$$

To reduce heavy-tail instability, we winsorize citation counts within each field-year at the 99th percentile before ranking.

### D.5.2. VENUE-LEVEL STANDING PERCENTILE (TYPE-AWARE BUT UNIFIED OUTPUT)

We define a venue standing percentile  $V_{\text{norm}}(p) \in [0, 1]$  based on the venue type:

**Journals.** If  $p$  is published in a journal covered by JCR, we use the JCR percentile rank of its Journal Impact Factor within the journal's subject category:

$$V_{\text{norm}}(p) = J_{\text{norm}}(p) = \text{Rank}_{\%}(\text{IF}(p) \mid \text{JCR\_Field}(p)).$$

**Conferences / proceedings.** Conferences typically lack JCR impact factors, but they are indexed with venue-level standing signals (e.g., proceedings series metrics, venue citation rates). We compute a robust proxy in two stages:

$$V_{\text{norm}}(p) = \text{Rank}_{\%}(M_{\text{conf}}(v) \mid \text{Field}(p)),$$

where  $v$  is the conference venue (or proceedings series) and  $M_{\text{conf}}(v)$  is defined by the best available signal in descending priority:

$$M_{\text{conf}}(v) = \begin{cases} \text{venue metric percentile from an index (e.g., CiteScore/SJR) if available,} \\ \text{two-year venue citation rate } \frac{\text{Cite2Y}(v)}{\text{Works2Y}(v)} \text{ from the same index,} \\ \text{fallback: long-run venue citation rate } \frac{\text{CiteAll}(v)}{\text{WorksAll}(v)}. \end{cases}$$

This design anchors conference standing in *the indexing institution's venue-level ranking signals* when present, and otherwise in a conservative, field-normalized citation-rate proxy that is stable under sparsity.

**Preprints.** Preprints are not peer-reviewed venues, yet they can be influential. We therefore compute a repository-level standing percentile and apply a conservative discount to reflect the absence of formal peer review:

$$V_{\text{norm}}(p) = \rho_{\text{pre}} \cdot \text{Rank}_{\%}(M_{\text{repo}}(r) \mid \text{Field}(p)),$$

where  $r$  is the preprint repository (e.g., arXiv/bioRxiv/medRxiv) and  $M_{\text{repo}}(r)$  is computed analogously to conferences via repository citation rate proxies. We set  $\rho_{\text{pre}} = 0.85$  to prevent preprints from receiving inflated venue credit solely due to rapid diffusion, while still allowing highly cited preprints to contribute meaningfully through  $C_{\text{norm}}(p)$ .

**Unified influence score.** We combine paper-level and venue-level percentiles with fixed weights:

$$\mathcal{I}(p) = w_c \cdot C_{\text{norm}}(p) + w_v \cdot V_{\text{norm}}(p), \quad w_c = 0.6, w_v = 0.4.$$

As in the main text, the higher weight on  $C_{\text{norm}}$  mitigates venue “halo effects” and preserves strong signals from influential papers in niche venues or emerging areas.

#### D.5.3. EVIDENCE STATEMENT CREDIBILITY WEIGHTS

For each evidence statement  $e_j$  (cluster  $G_j$ ), let  $P_j$  be the set of *unique* witness papers that contributed at least one claim to that cluster. We define:

$$\text{Support}(e_j) = \sum_{p \in P_j} \mathcal{I}(p), \quad \gamma_j = \frac{\text{Support}(e_j)}{\sum_{i=1}^k \text{Support}(e_i)}.$$

This paper-level de-duplication ensures that repeated mentions within the same witness do not artificially inflate support.

### D.6. Relation Classification, Weighted Voting, and Confidence Calibration

Given (i) the citing-side paraphrase  $c_A$  and (ii) a set of canonical evidence statements  $E = \{e_1, \dots, e_k\}$  with credibility weights  $\{\gamma_j\}$ , ICSV evaluates how each  $e_j$  relates to  $c_A$  and then aggregates via a reliability-aware vote.

#### D.6.1. RELATION CLASSIFICATION PROMPT

##### System prompt (relation classification).

You are an expert scientific fact-checker.

Inputs:

- (1) Claim  $c_A$ : what paper A attributes to paper B (a claim-preserving paraphrase).
- (2) Evidence  $e_j$ : a canonical statement of what multiple other papers attribute to B about one aspect.

Task: decide the logical relation of  $e_j$  to  $c_A$ .

Labels:

- ENTAILS:  $e_j$  clearly supports  $c_A$  under the same scope/conditions.
- CONTRADICTS:  $e_j$  clearly conflicts with  $c_A$  (opposite finding, incompatible scope, or mutually exclusive conditions).
- NEUTRAL:  $e_j$  is about a different aspect, or is insufficient to judge  $c_A$ .

Rules:

- Be strict about scope and qualifiers. If scopes differ, prefer NEUTRAL unless the mismatch itself implies contradiction.
- Do NOT assume unstated details.

Return JSON only.

##### User prompt.

Claim  $c_A$  (about paper B):  
 $\{c_A\}$

Evidence  $e_j$  (community-attributed statement about paper B):  
 $\{e_j\}$

Return JSON:

```
{
  "label": "ENTAILS|CONTRADICTS|NEUTRAL",
  "justification": "one sentence, must cite the key scope/qualifier alignment or mismatch"
}
```

We run relation classification with deterministic decoding (temperature 0). To measure robustness, we additionally run a small self-consistency check (three independent decoding seeds at  $T = 0$ ; identical output is expected, but disagreements can occur due to parsing ambiguities). Let  $a_j \in [0, 1]$  denote agreement, computed as the fraction of runs that match the majority label.

We map labels to scalar votes  $v_j \in \{+1, 0, -1\}$  for ENTAILS/NEUTRAL/CONTRADICTS.

#### D.6.2. WEIGHTED CONSENSUS SCORE

We compute the core consensus score as in the main text:

$$\mathcal{V}_{\text{final}} = \sum_{j=1}^k v_j \cdot \gamma_j \in [-1, 1],$$

and apply thresholds  $T_{\text{support}} = 0.3$  and  $T_{\text{misdite}} = -0.3$ :

$$\text{Verdict} = \begin{cases} \text{Supported}, & \mathcal{V}_{\text{final}} > 0.3, \\ \text{Misdite}, & \mathcal{V}_{\text{final}} < -0.3, \\ \text{Undecidable}, & \text{otherwise}. \end{cases}$$

#### D.6.3. CALIBRATED CONFIDENCE SCORE (ROBUST UNDER DISAGREEMENT AND CONCENTRATION)

While  $|\mathcal{V}_{\text{final}}|$  is a useful base signal, it can be misleading when (i) the committee is small, (ii) weights are dominated by a single witness cluster, or (iii) evidence statements disagree. We therefore compute an adjusted confidence  $\text{Conf} \in [0, 1]$  that is conservative under these failure modes.

**Effective evidence size.** We use an effective number of evidence statements (weight diversity):

$$n_{\text{eff}} = \frac{1}{\sum_{j=1}^k \gamma_j^2},$$

which decreases when one cluster dominates.

**Weighted disagreement.** Let  $w_\ell = \sum_{j: R_j = \ell} \gamma_j$  be the total credibility mass assigned to relation label  $\ell \in \{\text{ENTAILS}, \text{NEUTRAL}, \text{CONTRADICTS}\}$ . Define normalized entropy:

$$H = -\frac{\sum_\ell w_\ell \log(w_\ell + \epsilon)}{\log 3},$$

with a small  $\epsilon$  for numerical stability.  $H$  increases with disagreement.

**Stability factor.** Let  $\bar{a} = \sum_{j=1}^k \gamma_j a_j$  be the credibility-weighted relation stability under self-consistency.

**Final confidence.** We define:

$$\text{Conf} = \underbrace{|\mathcal{V}_{\text{final}}|}_{\text{margin}} \cdot \underbrace{\min\left(1, \frac{n_{\text{eff}}}{K_{\min}}\right)}_{\text{evidence sufficiency}} \cdot \underbrace{(1 - H)}_{\text{disagreement penalty}} \cdot \underbrace{\bar{a}}_{\text{classification stability}}.$$

This confidence is high only when the committee is sufficiently large/diverse, the evidence mass coheres, and relation labels are stable.

#### D.7. Reliability-Aware Abstention and Diagnostic Reporting

ICSV is designed for high-stakes integrity workflows, where a false accusation is more harmful than an abstention. We therefore abstain whenever community evidence is not strong enough to support a traceable verdict.

**Abstention triggers.** ICSV returns UNDECIDABLE if any of the following holds:

- (i) **Insufficient witnesses:**  $|C_{\text{open}}| < K_{\min}$ , with  $K_{\min} = 6$ .
- (ii) **Low consensus margin:**  $\mathcal{V}_{\text{final}} \in [T_{\text{misdite}}, T_{\text{support}}] = [-0.3, 0.3]$ .
- (iii) **Low calibrated confidence:**  $\text{Conf} < 0.5$  (conservative default).
- (iv) **High disagreement:**  $H > 0.6$  (evidence splits across entail/contradict/neutral).

**What abstention means (and what it does not).** An abstention is not a “failure” mode; it is an explicit integrity constraint. Under paywall conditions, deep semantic verification can be underdetermined. ICSV therefore refuses to overreach when the committee cannot reliably reconstruct  $B$ ’s contribution with sufficient consensus.

**Auditable output bundle.** For every verdict (including UNDECIDABLE), ICSV outputs:

- (1)  $c_A$  and the final window radius used to stabilize it;
- (2) the witness set size  $|C_{\text{open}}|$  and each witness paper’s  $\mathcal{I}(p)$ ;
- (3) clusters  $\{G_j\}$ , evidence statements  $\{e_j\}$ , and their credibility weights  $\{\gamma_j\}$ ;
- (4) per-cluster relation labels  $R_j$ , votes  $v_j$ , and stability  $a_j$ ;
- (5)  $\mathcal{V}_{\text{final}}$ , verdict thresholds, and calibrated Conf with disagreement statistics.

This reporting makes each paywall decision traceable to specific community attributions and clearly communicates when evidence is insufficient or contradictory.

## E. Evaluation Details: Grading, Decoding, and Token Accounting

### E.1. Acc-pass@3: Definition and Grading Pipeline

**Candidate generation.** For each miscitation instance and each evaluated method, we draw  $K = 3$  independent samples from the underlying model by varying the random seed while keeping the decoding configuration fixed for that method. Each sample consists of: (i) a predicted validity label  $\in \{\text{SUPPORTED}, \text{MISCIATION}\}$ , and (ii) a free-form natural-language explanation of the decision. Any explicit abstention (e.g., UNDECIDABLE) is treated as an ordinary label for grading purposes and will be marked incorrect whenever it disagrees with the gold label or fails to provide a faithful rationale.

**Independent grader LLM.** We use an independent grader based on gpt-4o-2024-08-06 to decide whether a given sample is *fully correct*, jointly considering the label and the explanation. The grader sees the gold label and gold explanation together with a single model prediction and must output exactly one token: CORRECT or INCORRECT. The system and user prompts are:

#### System prompt (grader).

You are an expert scientific editor.

You will be given:

1. A gold label and gold explanation for a miscitation instance.
2. A model’s predicted label and predicted explanation for the same instance.

Your task is to decide whether the model’s prediction is FULLY CORRECT.

A prediction is FULLY CORRECT only if BOTH of the following hold:

- The predicted label exactly matches the gold label.
- The predicted explanation identifies the same underlying error mechanism as the gold explanation (not just a generic restatement of the label).

Be strict. If the explanation is vague, only partially matches the gold rationale, introduces incorrect reasoning, or misses key aspects of the gold explanation, you must treat the prediction as INCORRECT.

Respond with a single token: either CORRECT or INCORRECT.

Do not output any other text.

#### User prompt (grader).

[INSTANCE]

Citing context:

{CITING\_TEXT}

Gold label:  
{GOLD\_LABEL}

Gold explanation:  
{GOLD\_EXPLANATION}

[PREDICTION]  
Predicted label:  
{PRED\_LABEL}

Predicted explanation:  
{PRED\_EXPLANATION}

Respond with exactly one token:  
- CORRECT  
- INCORRECT

For each instance  $i$  and sample  $k \in \{1, 2, 3\}$ , the grader returns a verdict  $g_{ik} \in \{\text{CORRECT}, \text{INCORRECT}\}$ .

**Metric definition.** Let  $N$  denote the number of evaluation instances for a given regime (MisciteBench-Open or MisciteBench-Paywall). Acc-pass@3 for a method is defined as:

$$\text{Acc-pass@3} = \frac{1}{N} \sum_{i=1}^N \mathbf{1} \left[ \exists k \leq 3 \text{ such that } g_{ik} = \text{CORRECT} \right], \quad (15)$$

where  $\mathbf{1}[\cdot]$  is the indicator function. In words, an instance contributes 1 to the numerator if *at least one* of the three sampled predictions is graded as fully correct with respect to both the label and the explanation; otherwise it contributes 0. Predictions graded as INCORRECT for any reason (wrong label, incomplete rationale, hallucinated rationale, or abstention) are not counted.

**Human validation of the grader.** To assess the reliability of the automatic grader, human experts manually inspected more than 3,000 randomly sampled grader decisions drawn across models, regimes, and label types. Disagreement between human judgment and the grader was observed in fewer than 0.5% of cases, predominantly on borderline instances where the predicted explanation partially overlapped with the gold rationale but omitted some secondary details. We retain the grader’s strictness in these edge cases, as Acc-pass@3 is intended to measure *diagnostically faithful* explanations rather than loose paraphrases of the label.

## E.2. Decoding and Sampling Settings

Unless otherwise specified in the corresponding module appendix, we use the following decoding configurations:

**Deterministic components.** For single-step classification or scoring modules that should be deterministic, we use greedy decoding with temperature  $T = 0$  and the provider’s default nucleus and top- $k$  settings. This includes: (i) the NLI model used in ACSV Phase III, (ii) the relation classification and other local decision steps inside ICSV where no self-consistency is applied, and (iii) lightweight utility calls such as format validation and consistency checks.

**Self-consistency components.** For modules that rely on multi-step reasoning and benefit from diversity, we use small-ensemble self-consistency. Concretely, we sample  $M = 5$  completions with temperature  $T = 0.7$  and top- $p = 0.95$ , and then aggregate by majority vote:

- ACSV Phase IV (LRM deep reasoning for ambiguous cases).
- Taxonomy-aligned miscitation classification (Section 3.4).

If there is no strict majority, we fall back to the most frequent non-abstaining class; when the distribution is too diffuse or inconsistent, the module may emit an explicit `UNDECIDABLE` label, which is treated as incorrect under Acc-pass@3.

**Baseline prompts.** For the Full-Text and Search baselines, we use a mildly stochastic but low-variance configuration: temperature  $T = 0.2$  with the provider’s default top- $p$  and other decoding parameters. This allows limited exploration across the  $K = 3$  samples while keeping predictions reasonably stable for grading.

**Averaging across runs.** To reduce variance from global random seeds and any stochasticity in external services (e.g., web search for the Search baselines), we repeat each full evaluation three times with different random seeds. All reported numbers in the main text and appendix tables are the arithmetic mean over these three runs.

### E.3. Token Accounting and Token Economy

**Per-instance token counting.** For every method and evaluation instance, we record the total number of tokens consumed by the verification pipeline, including both inputs and outputs, and including all internal calls (retrieval, NLI, committee reasoning, taxonomy classification, etc.). Concretely, for each model call we log: (i) the number of input tokens, (ii) the number of output tokens, and sum these over all calls involved in processing that instance. We use the official tokenization for each provider (e.g., the same tokenizer used for billing) to avoid discrepancies between counted and billed tokens.

**Token Economy definition.** For a given backbone model, let  $\text{Tok}_{\text{FT}}$  denote the mean number of tokens per instance for the Full-Text baseline, and let  $\text{Tok}_{\text{BIBAGENT}}$  denote the mean number for BIBAGENT (specifically, ACSV in MisciteBench-Open) evaluated on the same set of instances. We define Token Economy as

$$\text{TokenEcon} = 1 - \frac{\text{Tok}_{\text{BIBAGENT}}}{\text{Tok}_{\text{FT}}}, \quad (16)$$

which can be interpreted as the fraction of tokens saved by BIBAGENT relative to the Full-Text baseline for that backbone. A value of  $\text{TokenEcon} = 0.794$ , for example, means that BIBAGENT uses 79.4% fewer tokens on average than the corresponding Full-Text setup.

To make this comparison meaningful, Token Economy is computed only on the subset of instances for which *both* methods return a non-abstaining verdict (i.e., they emit a concrete `SUPPORTED` or `MISCITATION` label). Instances on which one method abstains and the other does not are excluded from the token-economy calculation but are still included in Acc-pass@3 with abstentions treated as incorrect predictions.

## F. Evidence Committee Reliability Ablation

ICSV (Appendix D) is only useful in the paywalled regime if its *Evidence Committee* behaves in a predictable, conservatively calibrated way: when community evidence is rich, it should speak with high confidence and high precision; when evidence is thin or inconsistent, it should abstain rather than speculate. The main paper states that we observe a sharp reliability transition once an aspect of the paywalled source is supported by at least six independent witnesses. This appendix provides the quantitative ablation that underpins that claim and justifies the global committee-size threshold used by ICSV’s reliability-aware abstention rule.

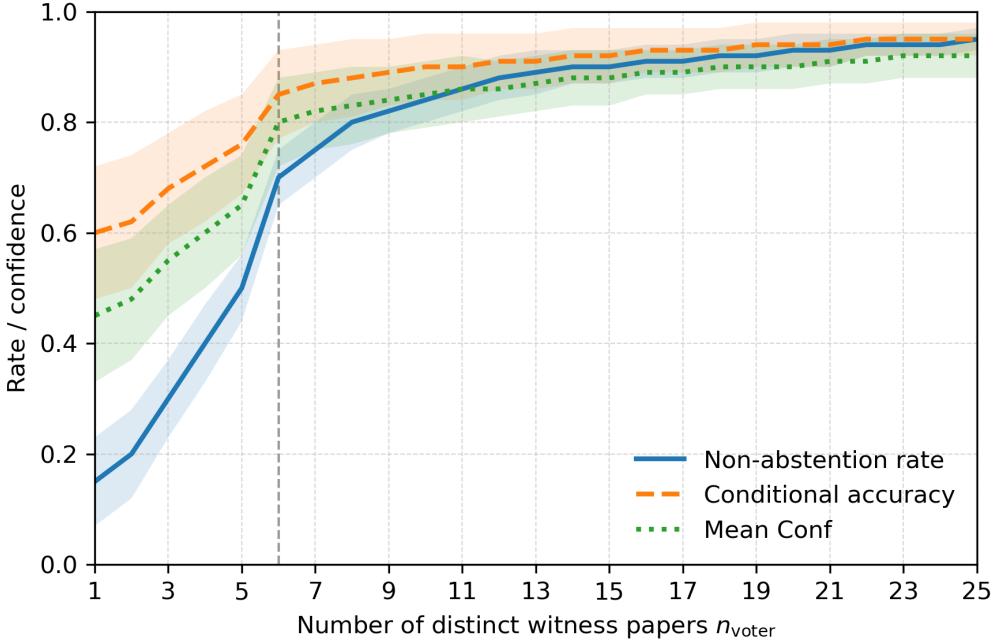
### F.1. Protocol

For a paywalled source  $B$ , ICSV groups all downstream open-access citers into semantic clusters  $G_1, \dots, G_k$ , where each  $G_j$  represents one coherent aspect of what the community attributes to  $B$  (method, dataset, empirical finding, etc.). Let  $P_j$  denote the set of *distinct* witness papers whose claims end up in  $G_j$ , and

$$n_j = |P_j|$$

the number of independent committee voters for that aspect. From each cluster we distill an evidence statement  $e_j$  with credibility weight  $\gamma_j$  (Appendix D), and define the *dominant* aspect for the citation as

$$j^* = \arg \max_j \gamma_j, \quad e^* = e_{j^*}.$$



*Figure 2.* Evidence Committee behavior as a function of the number of distinct witness papers  $n_{\text{voter}}$  supporting the dominant evidence statement  $e^*$  for a paywalled citation. Curves show non-abstention rate, conditional accuracy, and mean calibrated confidence (Appendix D); shaded bands indicate variation across backbones. The sharp and stable transition around  $n_{\text{voter}} = 6$  motivates the choice  $K_{\min} = 6$  in ICSV’s reliability-aware abstention rule.

Our ablation focuses on

$$n_{\text{voter}} = n_{j^*},$$

the number of distinct witness papers that support the aspect of  $B$  which most strongly drives ICSV’s verdict on that citation.

We run the unmodified ICSV pipeline on the full MisciteBench-Paywall split. For every paywalled citation that admits at least one non-empty cluster, and for each backbone, we record:

- the final verdict (*Supported*, *Miscitation*, or *Undecidable*);
- whether the verdict is correct under the benchmark label;
- the calibrated confidence score  $\text{Conf} \in [0, 1]$  from Appendix D, which combines the consensus margin  $|\mathcal{V}_{\text{final}}|$ , effective committee size, label disagreement, and self-consistency stability.

We then bucket citations by the integer value of  $n_{\text{voter}}$  in the range  $1 \leq n_{\text{voter}} \leq 25$ . For each bucket  $c$ , and for each backbone, we compute:

1. the *non-abstention rate*: the fraction of citations with  $n_{\text{voter}} = c$  on which ICSV outputs *Supported* or *Miscitation* rather than *Undecidable*;
2. the *conditional accuracy*: the fraction of those non-abstaining verdicts that match the ground truth;
3. the mean calibrated confidence  $\mathbb{E}[\text{Conf} \mid n_{\text{voter}} = c]$ .

The curves in Figure 2 plot these quantities after averaging over backbones; shaded bands indicate the range across backbones. All hyperparameters and thresholds are identical to those used in the main experiments; we do not re-tune ICSV for this study.

## F.2. Quantitative Trends

Three regimes emerge consistently across models and disciplines.

**Single or few witnesses ( $n_{\text{voter}} \leq 2$ ).** When the dominant aspect of  $B$  is supported by only one or two downstream papers, ICSV behaves cautiously. Non-abstention rates are low, and the system frequently returns *Undecidable* because the consensus margin and effective committee size terms in Conf are small. Among the few cases where ICSV does commit, conditional accuracy is noticeably weaker and confidence scores cluster in a moderate band. In this regime, individual witness papers often describe  $B$  in idiosyncratic or overly generic ways, and a single outlier can heavily influence the vote. The abstention mechanism therefore activates often, which is precisely the desired behavior for a conservative verifier operating under sparse community evidence.

**Small committees ( $3 \leq n_{\text{voter}} \leq 5$ ).** As  $n_{\text{voter}}$  grows into the range of three to five independent witnesses, non-abstention rates and conditional accuracy both improve: the committee has enough redundancy to filter out extreme outliers and resolve many straightforward paywalled cases. However, the curves in Figure 2 still exhibit noticeable variability across buckets in this regime. The calibrated confidence Conf rises compared to the  $n_{\text{voter}} \leq 2$  regime, but remains in an intermediate range, reflecting two residual sources of uncertainty: (i) disagreement between clusters about the precise scope of  $B$ 's contribution, and (ii) moderate label entropy when witness papers emphasize different aspects of  $B$  or mix descriptive and evaluative citations. In practice, ICSV continues to abstain on a substantial fraction of citations here, and the system remains deliberately conservative.

**Reliability transition ( $n_{\text{voter}} \geq 6$ ).** The key phenomenon appears once the dominant aspect is supported by at least six distinct witness papers. Beyond this point, all three metrics undergo a clear and stable shift:

- The non-abstention rate rises sharply and then plateaus: ICSV now produces concrete *Supported* or *Miscitation* verdicts on the majority of paywalled citations in these buckets, because both the consensus margin and the effective committee size become large enough to pass the internal reliability checks in Appendix D.
- Conditional accuracy of non-abstaining verdicts increases and stabilizes at a high level. Across backbones, once  $n_{\text{voter}} \geq 6$ , the empirical precision of ICSV's paywalled decisions is consistently high, and additional witnesses beyond six yield only marginal gains.
- The mean calibrated confidence  $\mathbb{E}[\text{Conf} | n_{\text{voter}} = c]$  crosses 0.8 and remains in a high-confidence band for all  $c \geq 6$ . In other words, whenever a paywalled aspect is supported by six or more independent witnesses, ICSV not only decides more often, but does so with confidence scores that reflect genuinely strong and internally coherent community evidence.

Importantly, this transition is not an artifact of a particular backbone or field. The same qualitative knee in the curves appears when the ablation is recomputed separately for each backbone and for coarse discipline groups (e.g., Clinical Medicine vs. Computer Science). The exact numeric values vary, but the location of the reliability transition—around six independent witnesses for the dominant aspect—is remarkably stable.

## F.3. Choice of Threshold and Robustness

The global threshold  $K_{\min} = 6$  used by ICSV's reliability-aware abstention rule is therefore not a hand-tuned hyperparameter, but the empirical knee point of a three-way trade-off:

- Below six witnesses, the Evidence Committee is too small to be trustworthy: non-abstention rates are lower, conditional accuracy is noticeably weaker, and the calibrated confidence Conf correctly reflects this by staying in a cautious range. Allowing aggressive decisions here would yield an undesirable increase in false accusations under genuine information scarcity.
- At six witnesses, the curves in Figure 2 enter a stable high-precision regime. Both the committee size and the consensus structure are sufficient for ICSV to exploit the community's distributed memory of the paywalled source, and the confidence calibration in Appendix D rewards this with high Conf values.

- Raising  $K_{\min}$  beyond six would sacrifice coverage for little additional precision. While larger committees remain slightly more stable, the incremental gain is small compared to the loss in the number of paywalled citations that can be decided at all, especially in niche subfields where only a handful of downstream citers exist.

We further stress-test this choice by recomputing the ablation under several perturbations: varying the confidence threshold used to declare a verdict vs. abstention, subsampling the witness set, and repeating the analysis on random halves of MsciteBench-Paywall. In all cases, the location of the knee in the reliability curves remains close to six witnesses, even though the absolute values of the metrics shift slightly. This robustness suggests that  $K_{\min} = 6$  captures a property of the underlying citation graph—when the literature around a paywalled source is broad and coherent enough to support stable reconstruction—rather than an artifact of a particular model or parameter setting.

Taken together, these results substantiate the design of ICSV’s Evidence Committee. The system only “speaks with conviction” about paywalled sources when the dominant aspect is backed by a sufficiently large and internally consistent community of citers, and it is willing to abstain explicitly when that condition is not met. This calibrated behavior is essential for deploying BIBAGENT in high-stakes editorial and auditing workflows: it ensures that paywall robustness is grounded in measurable community redundancy rather than in unexamined model confidence.