

BibTeX Citation Errors in Scientific Publishing Agents: Evaluation and Mitigation

Delip Rao*

University of Pennsylvania
delip@seas.upenn.edu

Chris Callison-Burch

University of Pennsylvania
ccb@seas.upenn.edu

Abstract

Large language models with web search are increasingly used in scientific publishing agents, yet they produce BibTeX entries with pervasive field-level errors stemming from omission, partial corruption, substitution, and hallucination. We construct a benchmark of 931 papers across four domains and three citation tiers—popular, low-citation, and recent post-cutoff—with version-aware ground truth. Three search-enabled frontier models (GPT-5, Claude Sonnet 4.6, Gemini 3 Flash) generate approximately 23,000 field-level observations. Overall accuracy is 83.6%, but only 50.9% of entries are fully correct; accuracy drops 27.7 pp from popular to recent papers, revealing heavy reliance on parametric memory even when search is available. Co-occurrence analysis identifies two failure modes: wholesale entry substitution and isolated field error. We present `clibib`, an open-source tool for deterministic BibTeX retrieval, as a mitigation mechanism. Two-stage integration raises accuracy to 91.5% (+8.0 pp) and fully correct entries to 78.3%, with a 0.8% regression rate. Separating search from revision yields larger gains and lower regression than single-stage tool loops (0.8% vs. 4.8%), demonstrating that integration architecture matters independently of model/tool capability. We release `clibib` and an accompanying agent skill under the MIT License to improve citation accuracy in increasingly automated scientific workflows.

1 Introduction

Bibliographic metadata is one of the few LLM outputs verifiable against authoritative records: every field in a BibTeX entry has a deterministic correct value in publisher databases. When models fabricate or corrupt these fields, errors propagate into manuscripts and downstream citation analyses. Prior work reports hallucination rates of 14–95% in LLM-generated citations (Walters & Wilder, 2023; Chelli et al., 2024; Aljamaan et al., 2024; Xu et al., 2026), but these studies evaluate base models without web search or rely on manual verification. Researchers now use frontier models with native search tools, yet no study systematically evaluates BibTeX accuracy under these conditions.

LLM-based agents now automate scientific workflows from literature review to manuscript preparation (Gridach et al., 2025; Schmidgall et al., 2025). We address this gap with a benchmark spanning four domains and three citation tiers—popular, low-citation, and recent post-cutoff—designed to disentangle parametric memory from search dependence. If search tools fully compensated for knowledge gaps, accuracy would be uniform across tiers; instead, we find a 27.7 pp drop from popular to recent papers. Field-error co-occurrence analysis identifies two distinct failure modes: wholesale entry substitution (identity fields fail together) and isolated field error.

We develop and evaluate `clibib`, an open-source tool that fetches authoritative BibTeX from the Zotero Translation Server via deterministic identifier resolution. Comparing single-stage tool loops with two-stage pipelines that separate search from revision, we find that the

*Corresponding author

latter yields larger gains with lower regression, demonstrating that integration architecture matters independently of model capability.

Our contributions: (1) a **benchmark** of 931 papers across four domains and three citation tiers with version-aware ground truth; (2) a **systematic evaluation** of three search-enabled frontier models producing 23,000 field-level observations on a six-way error taxonomy (Section 4); and (3) an open-source tool, **clibib**, evaluated as mitigation—two-stage integration raises fully correct entries from 50.9% to 78.3% with 0.8% regression (Section 5).

2 Related Work

Citation hallucination includes fabricated references and corrupted bibliographic metadata. Research measures fabrication rates and field-level accuracy and develops mitigation tools. We give a brief overview here; Appendix A provides a comprehensive survey.

LLMs fabricate plausible but fictitious references at rates ranging from 18% to 95% depending on the model and elicitation method (Alkaisi & McFarlane, 2023; Walters & Wilder, 2023; Xu et al., 2026). Among real citations, title is consistently the most accurate field while DOI is the worst; numerical metadata (volume, pages) are error-prone across all models (Chelli et al., 2024; Szeider, 2026). Niimi (2025) found a log-linear correlation between citation count and accuracy, suggesting a memorization threshold tied to training-data redundancy. Ansari (2026) showed that hallucinated citations in NeurIPS 2025 papers universally exhibit compound failures across multiple fields.

Cross-study comparison is hampered by methodological fragmentation: studies use different elicitation strategies, field sets (3 to 7+), and hallucination definitions (binary, threshold-based, weighted composites, categorical) (Walters & Wilder, 2023; Chelli et al., 2024; Aljamaan et al., 2024; Szeider, 2026). No standardized benchmark existed before CiteAudit (Yuan et al., 2026), which itself checks only six fields.

Mitigation approaches converge on bypassing LLM parametric memory through retrieval-augmented (Lewin et al., 2021) or tool-augmented (Nakano et al., 2022) generation. Systems such as CiteAudit, CheckIfExist (Abbonato, 2026), and BibAgent (Li et al., 2026) ground citations in authoritative databases, consistently reducing both fabrication and field-level errors. Our clibib builds on this insight by fetching authoritative BibTeX entries from the Zotero Translation Server.

3 Methodology

We prompt three frontier LLMs with natural-language paper descriptions and compare generated BibTeX against multi-source ground truth.

3.1 Task formulation

We adopt a *known-item retrieval* elicitation strategy with natural-language descriptions (KR-nat). Each prompt provides a paper’s title and first author (e.g., “Title” by First Author et al.), deliberately omitting DOI, year, venue, and volume to mirror how researchers naturally request citations. This contrasts with open-ended generation (Walters & Wilder, 2023; Chelli et al., 2024) and the degraded descriptions in Szeider (2026). The interaction is single-turn; responses are extracted via XML-style <bibtex> tags. Each of the 931 papers is queried once per model, yielding 2,793 generated entries.

3.2 Models and search configuration

We evaluate three frontier LLMs with vendor-provided web search enabled (Table 1): GPT-5 (OpenAI), Claude Sonnet 4.6 (Anthropic), and Gemini 3 Flash (Google), covering the range of search architectures currently available. We use native vendor APIs; generation uses no system prompt.

Table 1: Model configurations for BibTeX generation. All models use single-turn prompting with search tools force-enabled and a maximum output length of 8,192 tokens. GPT-5 does not expose a temperature parameter; Claude Sonnet 4.6 and Gemini 3 Flash use temperature 1.0.

Model	API model ID	Search mechanism	Cutoff
GPT-5	gpt-5-2025-08-07	web_search (Responses API)	Oct 2024
Claude Sonnet 4.6	claude-sonnet-4-6	web_search_20250305 (beta)	Aug 2025
Gemini 3 Flash	gemini-3-flash-preview	Google Search grounding	Jan 2025

Cutoff dates are vendor-reported knowledge cutoffs. Anthropic additionally reports a broader training data cutoff of January 2026 for Claude Sonnet 4.6; we report the stricter knowledge cutoff. These dates are self-reported and may not precisely bound what a model has seen.

Table 2: Paper statistics by domain and citation tier after quality filtering. Med. cit. is the median cited_by_count. DOI column reports coverage rate.

Domain	Tier	<i>n</i>	Med. cit.	Years	DOI
AI	Popular	96	1,437	2010–2022	70%
	Low-citation	90	1	2015–2024	82%
	Recent	49	22	2025–2026	100%
Medicine	Popular	100	3,567	2010–2022	100%
	Low-citation	93	0	2015–2024	100%
	Recent	45	0	2026	100%
MatSci	Popular	100	2,577	2010–2020	100%
	Low-citation	88	7	2015–2025	100%
	Recent	49	0	2026	100%
QuantComp	Popular	97	500	2010–2022	100%
	Low-citation	76	4	2015–2025	100%
	Recent	48	0	2026	100%
Total		931			95%

3.3 Paper selection

We select papers across four domains—AI, Medicine, Materials Science, and Quantum Computing—from 10 venues each (Appendix E, Table 7). Papers are sampled from the OpenAlex API and stratified into three citation tiers: **Popular** (100/domain): highest-cited, 2010–2022, testing parametric memory. **Low-citation** (100/domain): ≤ 25 th-percentile citation count, 2015–2025, requiring search. **Recent** (50/domain): published 2025–2026, post-cutoff, search-dependent. After ground truth construction and quality filtering (Section 3.4), the final dataset contains 931 papers (Table 2; Appendix E).

3.4 Ground truth construction

A single paper may have an arXiv preprint, proceedings, and journal version, each with legitimately different BibTeX. Our ground truth accounts for this version multiplicity: we discover all citable versions, fetch authoritative BibTeX for each, and cross-validate against domain-specific databases.

Version discovery. Each paper’s OpenAlex record lists hosting locations with source-type metadata. We classify locations as arXiv, proceedings, or journal using URL and source-type heuristics; mirrors (PubMed/PMC, repositories) are excluded. The 997 papers yield 1,131 fetchable versions (Appendix E, Table 8).

BibTeX retrieval. For each fetchable version, `clibib` resolves identifiers through the Zotero Translation Server, with title-match validation requiring ≥ 0.85 token overlap. Of 1,131

fetchable versions, 955 (84.4%) returned valid BibTeX; Medicine is lowest (49.4%) due to paywalled journals, but PubMed fills the gap for 148 of 250 medicine papers (Appendix E).

Cross-validation. We cross-validate clibib results against domain-specific authoritative databases—DBLP for AI and Quantum Computing, PubMed for Medicine—with match rates and field-level discrepancies detailed in Appendix E.

Canonical field resolution. We resolve one canonical value per field using source priority rules (DOI from OpenAlex, title from Zotero, author and venue from DBLP or PubMed with Zotero fallback, year by majority vote, volume/pages from the published version). Of 997 papers, 975 (97.8%) have at least one canonical field resolved (Appendix E, Figure 7).

Venue synonym table. A synonym table maps the 40 venues to 141 name variants (abbreviations, full names, naming conventions across sources). Of 2,793 venue evaluations, 99 (3.5%) were escalated to Stage 2; only 5 were formatting mismatches, with the remaining 94 reflecting genuine model errors (Appendix E).

Quality filtering. The dataset was reduced from 997 to 931 papers: 20 removed during ground truth reconstruction (metadata conflicts) and 46 during a quality audit (errata, off-topic items). All results report post-filtering counts.

3.5 Error taxonomy

We evaluate 9 fields per entry: `entry_type`, `author`, `title`, `year`, `venue`, `volume`, `number`, `pages`, and `doi` (`entry_key` is excluded as arbitrary). Each field receives one of six labels: **C** (correct), **M** (missing from generated entry), **F** (fabricated—no verifiable source), **P** (partially correct—meaningful overlap but inexact), **S** (substituted—from a different paper), or **X** (not applicable for this entry type). The F/S distinction matters for diagnosis: fabrication indicates generation from noise; substitution indicates retrieval of the wrong record.

3.6 Evaluation pipeline

The evaluation pipeline assigns labels to 25,137 field-level observations ($2,793 \text{ entries} \times 9 \text{ fields}$) in two automated stages, with a human spot-check.

Stage 1: Deterministic classification. Rule-based matching assigns C, M, or X labels using per-field normalization: author by first-author last name (case-insensitive), titles lowercased with LaTeX removed, venues resolved through the synonym table, DOIs lowercased with prefixes stripped, pages normalized to start–end format, years by exact match. A field is labeled C if it matches any ground-truth version. Stage 1 resolved 93.6% of evaluable fields (21,631 of 23,103).

Stage 2: LLM-as-judge. The remaining 1,472 fields are classified by Gemini 3.1 Flash Lite via the autorubric framework (Rao & Callison-Burch, 2026) using two binary criteria: *partial_match* (meaningful overlap?) and *different_paper* (from a different work?), mapped to P, S, or F labels. We acknowledge the vendor overlap with the evaluated Gemini 3 Flash, but three factors mitigate bias: Flash Lite is a distinct variant, human validation shows no systematic favoring ($\kappa = 0.67\text{--}0.84$), and Stage 2 covers only 6.4% of evaluable fields. Appendix B documents the full rubric.

Stage 3: Human spot-check. A stratified sample of 521 fields (15% of the Stage 2 pool) was independently re-annotated, yielding 80.4% agreement ($\kappa = 0.67$). Details are in Section 6.4.

4 Results

Across 23,103 evaluable fields, 83.6% are correct; only 50.9% of entries have all fields correct.

Table 3: Per-field accuracy (% C of evaluable fields) by model. Fields ordered by overall accuracy. Entry_key excluded (always X).

Field	GPT-5	Claude 4.6	Gemini 3 Fl.	All
entry_type	90.3	86.1	96.9	91.1
author	93.5	86.0	93.9	91.1
year	87.4	83.1	94.4	88.3
venue	84.8	83.4	94.7	87.6
volume	74.8	82.4	92.6	83.3
title	78.5	78.1	85.3	80.6
pages	75.0	75.9	81.7	77.5
doi	75.2	72.8	78.6	75.5
number	65.6	72.5	77.8	72.0
Overall	81.4	80.4	88.9	83.6

4.1 Overall error rates

Table 3 shows per-field accuracy. Author and entry type lead (91.1% each); number (72.0%) and DOI (75.5%) are least accurate, consistent with prior findings (Walters & Wilder, 2023; Chelli et al., 2024). Missing values dominate errors for number and DOI; partial matches concentrate in title (13.9%). Author matching uses first-author last name; stricter all-author matching reduces accuracy by 11.1 pp (Appendix F, Tables 11 and Figure 11).

4.2 Cross-model comparison

Gemini 3 Flash leads overall (88.9%), followed by GPT-5 (81.4%) and Claude Sonnet 4.6 (80.4%), with pairwise differences significant at $p < 0.001$ (Table 3; per-field breakdown in Appendix F, Figure 8). Despite an 8.5 pp field-level gap, Claude and Gemini achieve similar fully correct rates (53.0% vs. 53.6%); GPT-5 achieves 46.1%. Error profiles differ: GPT-5 has the highest fabrication rate (2.2%), Claude the highest missing rate (13.8%), and Gemini the lowest error rates across all categories (Appendix F, Figure 12).

4.3 Cross-domain variation

Domain effects are moderate. Materials science and quantum computing achieve the highest accuracy (86.8% each), followed by AI (83.1%). Medicine is hardest at 77.7%, reflecting higher missing-field rates in biomedical journal metadata. Gemini leads in all four domains (Appendix F, Figure 9).

4.4 Citation tier effects

Citation tier has the largest effect of any variable we examine (Figure 1). Popular papers achieve 92.7% accuracy, low-citation papers 82.0%, and recent (post-cutoff) papers 65.0%—a 27.7 pp popular-to-recent gap. Per-model gaps range from 9.7 pp (Gemini) to 38.7 pp (Claude); Gemini maintains 83.4% even on post-cutoff papers, consistent with stronger search grounding. The low-citation-to-recent drop is where search degrades most. Pages shows the largest per-field decline (−51.8 pp), followed by DOI (−37.8 pp) and volume (−32.6 pp); author is most robust (−17.5 pp). Full per-field and per-model breakdowns are in Appendix F, Tables 9 and 12.

4.5 Search invocation and telemetry

Search tools are invoked for 97.4% of entries overall; when invoked, accuracy is 84.2%. GPT-5 without search achieves 85.1% on 49 entries (predominantly popular papers recalled from parametric memory), while Gemini and Claude without search produce no parseable BibTeX. Search depth varies widely: GPT-5 retrieves a median of 50 sources per entry, Claude 10, and Gemini 8. For recent papers, all models search on >97% of entries, yet

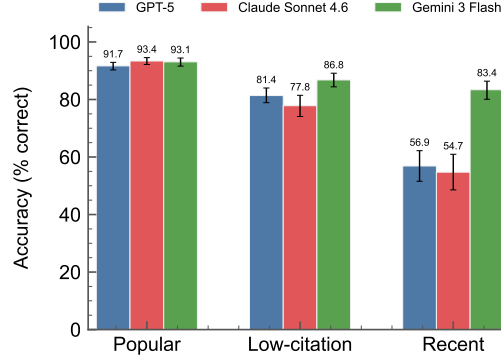


Figure 1: Field-level accuracy by citation tier and model with 95% bootstrap confidence intervals. Popular papers achieve 92.7%; recent (post-cutoff) papers drop to 65.0%.



Figure 2: Conditional co-error probability matrix. Cell (i, j) shows $P(\text{field } j \text{ is wrong} \mid \text{field } i \text{ is wrong})$. Entry type, author, and year errors strongly predict venue and DOI errors (0.74–0.91), consistent with wholesale entry substitution.

accuracy remains at 65%—the bottleneck is not whether models search but how effectively they use search results (Appendix F, Figure 10).

4.6 Field-error co-occurrence

The co-error matrix (Figure 2) is asymmetric, revealing two distinct modes. The first is *wholesale entry substitution*: identity fields (author, year, venue) form a tightly coupled error cluster—author errors predict year errors (0.74) and venue errors (0.77); year errors predict DOI errors (0.91)—signaling retrieval of the wrong paper entirely. Ansari (2026) report a similar compound failure pattern in NeurIPS 2025 papers.

The second mode is *isolated field error*: title errors often occur without author errors ($P = 0.33$), reflecting formatting differences rather than wrong-paper retrieval. Volume and number errors co-occur at high rates (0.55–0.78), consistent with jointly specified metadata.

These modes require different mitigation: wholesale substitution demands external database lookup, while isolated errors can often be resolved by normalization. Checking three identity fields (author, venue, year) suffices to detect the more damaging mode. Among 1,372 entries with errors, isolated errors dominate (65.2%); clibib corrects 57.9% of isolated-error entries versus 14.3% of wholesale substitutions (Appendix C).

Algorithm 1 Two-stage clibib integration

Require: Paper metadata p , baseline BibTeX b_{base} , model M
Ensure: Revised BibTeX b

```

1:  $q \leftarrow p.url \parallel p.doi \parallel p.title$  ▷ Priority order
2: if  $q$  is empty then
3:   return  $b_{\text{base}}$  ▷ No query available
4: end if
5:  $r \leftarrow \text{CLIBIBLOOKUP}(q)$ 
6: if  $r.status = \text{not\_found}$  then
7:   return  $b_{\text{base}}$ 
8: end if
9:  $J \leftarrow \text{JACCARDSIMILARITY}(q, r.title)$ 
10: if  $J < 0.3$  then
11:   return  $b_{\text{base}}$  ▷ Title mismatch – wrong paper
12: end if
13:  $prompt \leftarrow \text{REVISIONPROMPT}(b_{\text{base}}, r.bibtex)$ 
14:  $b \leftarrow \text{M.GENERATE}(prompt, tools = \emptyset)$  ▷ No tools, single turn
15: return  $\text{EXTRACTBIBTEX}(b)$ 
```

5 clibib as Mitigation

We evaluate clibib, an open-source tool that fetches verified BibTeX from the Zotero Translation Server (aggregating CrossRef, DBLP, PubMed), as a mitigation mechanism.

5.1 System description

clibib is a Python CLI tool and LLM agent skill¹ that accepts DOIs, arXiv IDs, ISBNs, PMIDs, titles, or URLs and routes each query to the appropriate Zotero Translation Server² endpoint. Identifier-based queries resolve deterministically—the server returns exactly one publisher-deposited record without LLM involvement. Free-text title queries trigger a search ranked by token-level Jaccard similarity, with a CrossRef API fallback. The tool ships on PyPI as an agentskills.io-compatible skill (~230 lines of Python). Deterministic resolution makes clibib effective for filling missing fields, but title-based search remains unreliable, and 9.5% of lookups returned `not_found` during ground truth construction. Implementation details are in Appendix G.

5.2 Mitigation protocol

We evaluate two integration architectures: a *single-stage* tool loop and a *two-stage* revision pipeline.

Single-stage (GPT-5 and Claude). The `clibib_lookup` function is registered alongside web search; the model drives a tool loop (up to 3 rounds) to find the paper’s identifier and call `clibib_lookup` (Algorithm 2, Appendix G). Gemini is excluded because its API does not allow search grounding and function declarations in the same request.

Two-stage (all models). Phase 1 uses each model’s baseline output. Phase 2 performs a client-side clibib lookup using the paper’s URL, DOI, or title, validated by a title-match gate (Jaccard ≥ 0.3). If the lookup succeeds, the model receives a revision prompt (Appendix D) for field-by-field reconciliation (Algorithm 1).

The revision prompt implements asymmetric reconciliation: matching titles trigger field replacement; mismatching titles preserve the baseline. All augmented entries are evaluated through the identical pipeline (Section 3.6).

¹<https://github.com/delip/clibib>

²<https://github.com/zotero/translation-server>

Table 4: Field-level accuracy (%) by model and integration strategy. Δ = gain over baseline (pp). Gemini lacks single-stage support.

Model	Baseline	Single	Δ	Two-st.	Δ
GPT-5	81.4	86.7	+5.3	92.5	+11.1
Claude Sonnet 4.6	80.4	84.0	+3.6	87.2	+6.8
Gemini 3 Flash	88.9	—		94.9	+6.0
All	83.6	—		91.5	+8.0

5.3 Mitigation results

Single-stage `clibib` augmentation (GPT-5 and Claude, 1,862 entries) raised aggregate accuracy from 80.9% to 85.3% (+4.4 pp), correcting 43.4% of baseline errors with a 4.8% regression rate (Appendix G, Figure 13). GPT-5 improved by +5.3 pp and Claude by +3.6 pp; fully correct entries rose from 46.1% to 69.2% (GPT-5) and 53.0% to 70.4% (Claude). The largest per-field gains are in structured numeric fields: number (+19.1 pp), volume (+9.0 pp), pages (+7.9 pp), and doi (+7.2 pp) (Appendix G, Figure 14). Missing fields had the highest correction rate (49.2%); fabricated and substituted errors resist correction because the lookup query inherits the model’s confusion about which paper was intended (Appendix G, Tables 13 and 14).

5.4 Ablation: single-stage vs. two-stage integration

The single-stage evaluation (Section 5.3) covers GPT-5 and Claude only (Gemini excluded per Section 5.2). To compare all three models under a uniform architecture, we apply a two-stage approach: each model’s baseline entry is revised against a client-side `clibib` lookup using the same revision prompt (Section 5.2). This also lets us compare the single-stage tool loop (GPT-5 and Claude) against the two-stage revision approach on the same models.

Table 4 compares the three conditions. Two-stage integration raises overall accuracy to 91.5% (+8.0 pp over baseline), correcting 52.3% of errors with a 0.8% regression rate; fully correct entries rise from 50.9% to 78.3%.

Two-stage outperforms single-stage for both GPT-5 (+5.8 pp) and Claude (+3.2 pp), with differences significant at $p < 10^{-4}$ (McNemar’s test). The two-stage approach also reduces regressions: 0.4–1.0% versus 4.3–5.3% for single-stage. The focused revision task avoids overloading the model with concurrent search and revision. When the strategies disagree (11.7% of fields), two-stage wins $2.4\times$ more often (Appendix G, Figures 16 and 17).

6 Discussion

6.1 Search-enabled models: better but not solved

Web search improves BibTeX generation to 83.6% field accuracy, above base-model evaluations (Walters & Wilder, 2023; Chelli et al., 2024; Xu et al., 2026), but fewer than half of entries are fully correct. The 27.7 pp popular-to-recent gap indicates heavy reliance on parametric memory: models invoked search on 97.4% of entries, yet a common pattern is locating the right paper but reconstructing BibTeX from memory.

6.2 Why `clibib` works and where it falls short

`clibib` is most effective for missing and partially correct fields (correction rates 49% and 48%), where the right paper has incomplete metadata. Fabricated and substituted fields resist correction because the lookup query inherits the model’s confusion about which paper was intended. The two-stage ablation confirms that separating retrieval from revision yields larger gains and lower regression than single-pass tool loops. Lookup coverage remains the

primary bottleneck, particularly for AI workshop papers and recent publications lacking indexed DOIs (Appendix G).

Ground-truth independence. Because clibib queries the same Zotero infrastructure used to construct ground truth, we test for circularity by restricting analysis to the three domains with independent cross-validation (DBLP for AI/Quantum Computing, PubMed for Medicine; 694 papers, 75% of the dataset). On this subset, the two-stage delta is *larger*: +9.8 pp field-level (vs. +8.0 pp overall). Materials science—the one domain without independent cross-validation—shows the *smallest* gains (+2.8 pp), the opposite of what circularity would predict.

6.3 Comparison with prior work

Our findings corroborate prior work: titles are most accurate, DOIs least, numerical fields error-prone. Search tools improve substantially over base models (75.5% DOI accuracy vs. Chelli et al. 2024’s 16–20%), and our tier stratification provides converging evidence for Niimi’s memorization-threshold hypothesis. Szeider (2026)’s unmediated DBLP export achieves 82.7% perfect entries versus our 50.9% baseline; two-stage clibib narrows this to 78.3% (Appendix H).

6.4 Human validation of LLM judge labels

A stratified sample of 521 fields (15% of the Stage 2 pool) was independently re-annotated by a human judge. Overall agreement is 80.4% with Cohen’s $\kappa = 0.67$; excluding two systematic disagreement patterns (entry_type classification and preprint-vs-published DOI versioning, which account for 66 of 102 disagreements) raises κ to 0.84. The direction is conservative: the judge over-labels F and under-labels P, so reported error rates are pessimistic. A sensitivity analysis confirms that disagreements shift aggregate accuracy by less than 0.2 pp (Appendix F, Table 10).

7 clibib as an Agent Skill

LLM workflows are moving from isolated chat interactions to agent harnesses that connect models to files, shells, and external tools (Yang et al., 2024). General-purpose examples include Pi, OpenCode, Codex, and Claude Code; scientific workflows are following the same pattern (Gridach et al., 2025; Schmidgall et al., 2025). To support ecologically valid deployment beyond our controlled pipelines, the accompanying MIT-licensed clibib release³ includes an agent skill that incorporates deterministic retrieval into bibliography management. In hosts that expose slash-command activation, users can invoke the skill as /clibib; citation-related requests can also select it contextually.

An agent can retrieve a new reference or orchestrate repeated lookups to repair a bibliography from prompts such as:

Adding a reference.

Add the reference for "Attention Is All You Need"

Repairing a bibliography.

Use the clibib skill to fix my .bib file

A persistent instruction in AGENTS.md or CLAUDE.md makes validation the default policy for new references:

Always use the clibib skill to validate any new references.

³<https://github.com/delip/clibib>

8 Conclusion

Search-enabled LLMs produce better BibTeX than base models, but half of entries contain at least one error. The dominant factor is paper recency: accuracy drops 27.7 pp from popular to post-cutoff papers despite near-universal search invocation. Co-occurrence analysis reveals that identity fields (author, venue, year) fail together in wholesale substitutions, while content fields fail independently; checking three identity fields suffices to detect the most damaging mode. Routing LLM output through deterministic retrieval produces substantial gains with minimal regression. Separating retrieval from revision (two-stage) outperforms single-pass tool loops, a lesson that likely generalizes to other settings where LLMs must incorporate structured external evidence. Open problems include coverage gaps for recent publications and the need for standardized BibTeX evaluation methodology.

Limitations

We test a single elicitation strategy (known-item retrieval with natural-language descriptions); open-ended generation may produce different error profiles. Our benchmark covers four scientific domains; fields with different publication norms (law, humanities) may behave differently. We evaluate three models at a single point in time. Ground truth coverage depends on OpenAlex, DBLP, PubMed, and the Zotero Translation Server; approximately 11% of `clibib` lookups returned `not_found`. Domain-level comparisons should note that ground-truth provenance varies: AI and Quantum Computing rely on DBLP and Zotero, Medicine on PubMed, Materials Science on Zotero alone. The circularity concern (shared Zotero infrastructure between `clibib` and ground truth) is addressed by the restricted-subset analysis in Section 6.2. The autorubric Stage 2 evaluation introduces error; human validation yields $\kappa = 0.67$ overall (Section 6.4). We use vendor-recommended temperatures; results at $T = 0$ may differ. Our primary author evaluation matches first-author last name only; stricter matching reduces accuracy by 11.1 pp (Appendix F, Table 11). We cover English-language venues only and do not compare against purpose-built tools such as `CheckIfExist`.

Acknowledgments

This research was developed with funding from the Defense Advanced Research Projects Agency’s (DARPA) SciFy program (Agreement No. HR00112520300) and is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via 56000026C0019 (the BENGAL program). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of DARPA, ODNI, IARPA, the Department of Defense, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

Ethics Statement

This study uses publicly available paper metadata; no human subjects were involved. All seed papers were drawn from public databases (OpenAlex, DBLP, PubMed). Detailed error taxonomies could theoretically help generate more convincing fake references, but the patterns are already widely known and our mitigation tool directly counteracts them.

Reproducibility Statement

The clibib software is independently available at <https://github.com/delip/clibib>; that repository does not contain the study benchmark or evaluation artifacts. The surviving study materials—prompt and rubric definitions, research code for sampling, generation, evaluation, and analysis, and rendered aggregate figures— are archived at <https://github.com/delip/colm-clibib-paper-supplement>.

Several weeks before preparation of the camera-ready version, an unrecoverable hardware failure resulted in the loss of the original 931-paper benchmark records, raw model outputs, complete field-level annotations, and underlying analysis tables before they could be archived. These artifacts therefore cannot be released. The surviving materials support inspection of the methods but not exact reproduction of the reported numerical results.

References

- Diletta Abbonato. CheckIfExist: Detecting Citation Hallucinations in the Era of AI-Generated Content, January 2026. URL <http://arxiv.org/abs/2602.15871>. arXiv:2602.15871.
- Ayush Agrawal, Mirac Suzgun, Lester Mackey, and Adam Tauman Kalai. Do Language Models Know When They're Hallucinating References?, March 2024. URL <http://arxiv.org/abs/2305.18248>. arXiv:2305.18248.
- Andres Algaba, Carmen Mazijn, Vincent Holst, Floriano Tori, Sylvia Wenmackers, and Vincent Ginis. Large Language Models Reflect Human Citation Patterns with a Heightened Citation Bias, August 2024. URL <http://arxiv.org/abs/2405.15739>. arXiv:2405.15739.
- Fadi Aljamaan, Mohamad-Hani Temsah, Ibraheem Altamimi, Ayman Al-Eyadhy, Amr Jamal, Khalid Alhasan, Tamer A Mesallam, Mohamed Farahat, and Khalid H Malki. Reference Hallucination Score for Medical Artificial Intelligence Chatbots: Development and Usability Study. *JMIR Medical Informatics*, 12:e54345, July 2024. ISSN 2291-9694. doi: 10.2196/54345. URL <https://medinform.jmir.org/2024/1/e54345>.
- Hussam Alkaissi and Samy I McFarlane. Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. *Cureus*, February 2023. doi: 10.7759/cureus.35179. URL <https://www.cureus.com/articles/138667-artificial-hallucinations-in-chatgpt-implications-in-scientific-writing>.
- Samar Ansari. Compound Deception in Elite Peer Review: A Failure Mode Taxonomy of 100 Fabricated Citations at NeurIPS 2025, February 2026. URL <http://arxiv.org/abs/2602.05930>. arXiv:2602.05930.
- Moses Boudourides. Structural Hallucination in Large Language Models: A Network-Based Evaluation of Knowledge Organization and Citation Integrity, 2026. URL <https://arxiv.org/abs/2603.01341>.
- Mikaël Chelli, Jules Descamps, Vincent Lavoué, Christophe Trojani, Michel Azar, Marcel Deckert, Jean-Luc Raynier, Gilles Clowez, Pascal Boileau, and Caroline Ruetsch-Chelli. Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis. *Journal of Medical Internet Research*, 26:e53164, May 2024. doi: 10.2196/53164. URL <https://www.jmir.org/2024/1/e53164>.
- I.-Chun Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, and Pengfei Liu. FacTool: Factuality Detection in Generative AI – A Tool Augmented Framework for Multi-Task and Multi-Domain Scenarios, July 2023. URL <http://arxiv.org/abs/2307.13528>. arXiv:2307.13528.
- Yee Man Choi, Xuehang Guo, Yi R. Fung, and Qingyun Wang. CiteGuard: Faithful Citation Attribution for LLMs via Retrieval-Augmented Validation, January 2026. URL <http://arxiv.org/abs/2510.17853>. arXiv:2510.17853.
- Mourad Gridach, Jay Nanavati, Christina Mack, Khaldoun Zine El Abidine, and Lenon Mendes. Agentic AI for Scientific Discovery: A Survey of Progress, Challenges, and Future Directions. In *Proceedings of the Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=TyCYakX9BD>.
- Mengze Hong, Di Jiang, Chen Jason Zhang, Zichang Guo, Yawen Li, Jun Chen, Shaobo Cui, and Zhiyang Su. CiteLLM: An Agentic Platform for Trustworthy Scientific Reference Discovery, February 2026. URL <http://arxiv.org/abs/2602.23075>. arXiv:2602.23075.
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions, November 2024. URL <http://arxiv.org/abs/2311.05232>. arXiv:2311.05232.

- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12):1–38, December 2023. doi: 10.1145/3571730. URL <https://dl.acm.org/doi/10.1145/3571730>.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, April 2021. URL <http://arxiv.org/abs/2005.11401>. arXiv:2005.11401.
- Peiran Li, Fangzhou Lin, Shuo Xing, Xiang Zheng, Xi Hong, Siyuan Yang, Jiashuo Sun, Zhengzhong Tu, and Chaoqun Ni. BibAgent: An Agentic Framework for Traceable Miscitation Detection in Scientific Literature, January 2026. URL <http://arxiv.org/abs/2601.16993>. arXiv:2601.16993.
- Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. WebGPT: Browser-assisted question-answering with human feedback, June 2022. URL <http://arxiv.org/abs/2112.09332>. arXiv:2112.09332.
- Junichiro Niimi. Hallucinations in Bibliographic Recommendation: Citation Frequency as a Proxy for Training Data Redundancy, 2025. URL <https://arxiv.org/abs/2510.25378>.
- Delip Rao and Chris Callison-Burch. Autorubric: A unifying framework for rubric-based LLM evaluation on non-verifiable tasks. In *Conference on Language Modeling (COLM)*, San Francisco, CA, October 2026.
- Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. HalluCitation Matters: Revealing the Impact of Hallucinated References with 300 Hallucinated Papers in ACL Conferences, January 2026. URL <http://arxiv.org/abs/2601.18724>. arXiv:2601.18724.
- Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Michael Moor, Zicheng Liu, and Emad Barsoum. Agent Laboratory: Using LLM Agents as Research Assistants. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 5977–6043, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.320. URL <https://aclanthology.org/2025.findings-emnlp.320>.
- Stefan Szeider. Unmediated AI-Assisted Scholarly Citations, February 2026. URL <http://arxiv.org/abs/2602.01686>. arXiv:2602.01686.
- William H. Walters and Esther Isabelle Wilder. Fabrication and errors in the bibliographic citations generated by ChatGPT. *Scientific Reports*, 13(1):14045, September 2023. doi: 10.1038/s41598-023-41032-5. URL <https://www.nature.com/articles/s41598-023-41032-5>.
- Zuyao Xu, Yuqi Qiu, Lu Sun, FaSheng Miao, Fubin Wu, Xinyi Wang, Xiang Li, Haozhe Lu, ZhengZe Zhang, Yuxin Hu, Jialu Li, Jin Luo, Feng Zhang, Rui Luo, Xinran Liu, Yingxian Li, and Jiaji Liu. GhostCite: A Large-Scale Analysis of Citation Validity in the Age of Large Language Models, February 2026. URL <http://arxiv.org/abs/2602.06718>. arXiv:2602.06718.
- John Yang, Carlos Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. SWE-agent: Agent-Computer Interfaces Enable Automated Software Engineering. In *Advances in Neural Information Processing Systems 37*, pp. 50528–50652, Vancouver, BC, Canada, 2024. Neural Information Processing Systems Foundation, Inc. (NeurIPS). ISBN 9798331314385. doi: 10.52202/079017-1601. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/5a7c947568c1b1328ccc5230172e1e7c-Abstract-Conference.html.
- Zhengqing Yuan, Kaiwen Shi, Zheyuan Zhang, Lichao Sun, Nitesh V. Chawla, and Yanfang Ye. CiteAudit: You Cited It, But Did You Read It? A Benchmark for Verifying Scientific References in the LLM Era, February 2026. URL <http://arxiv.org/abs/2602.23452>. arXiv:2602.23452.

Generative AI Use Disclosure

We used Gemini-Pro 3.0 and Claude Sonnet 4.5 for proofreading and plotting; all outputs were manually reviewed. We used `clibib` to create all BibTeX entries, manually reviewed for accuracy.

Outline of the appendix

The appendix is extensive, so we provide a brief roadmap to help reviewers locate specific material.

Appendix A *Extended related work.* Comprehensive survey of citation hallucination studies, field-level accuracy evaluations, and mitigation approaches—the full literature review condensed to a summary in Section 2.

Appendix B *LLM-as-judge rubric and borderline examples.* Full criteria definitions for the Stage 2 autorubric, the verdict mapping from binary criteria to P/S/F labels, few-shot calibration examples, and the two systematic disagreement patterns identified during human validation.

Appendix C *Failure mode examples.* Representative BibTeX entries illustrating wholesale entry substitution and isolated field error, with field-by-field comparison against ground truth.

Appendix D *Two-stage revision prompt.* The exact prompt template used for Phase 2 of the two-stage integration.

Appendix E *Dataset construction details.* Venue lists, dataset composition figures, version landscape statistics, BibTeX retrieval outcomes, cross-validation details, canonical field coverage, and venue synonym examples—supporting material for Section 3.

Appendix F *Supplementary evaluation results.* Per-field accuracy by model, domain accuracy, search telemetry, tier-level breakdowns, human agreement statistics, error-type distributions, author matching sensitivity analysis, and search depth details—supporting material for Section 4.

Appendix G *Supplementary mitigation results.* Single-stage algorithm pseudocode, per-field before/after comparisons, correction rates by error type, query routing details, lookup coverage, domain and tier effects, per-model correction patterns, strategy agreement analysis, and two-stage ablation figures—supporting material for Section 5.

Appendix H *Extended comparison with prior work.* Detailed per-study comparisons with [Walters & Wilder \(2023\)](#), [Chelli et al. \(2024\)](#), [Niimi \(2025\)](#), [Szeider \(2026\)](#), and [Boudourides \(2026\)](#)—expanding the summary in Section 6.3.

A Extended Related Work

A.1 Citation hallucination in LLM outputs

Hallucination in large language models—generating fluent text unsupported by training data or grounding sources—is well documented ([Ji et al., 2023](#); [Huang et al., 2024](#)). Citation hallucination is a specific instance: fabricated or corrupted references propagate into manuscripts and downstream citation analyses. When an LLM produces a bibliographic reference, two distinct failure modes arise: the model may fabricate a reference to a paper that does not exist, or it may cite a real paper with corrupted metadata (wrong authors, title, year, venue, DOI, or other fields). Both modes undermine the reliability of LLM-assisted scientific writing.

Early documentation of the problem focused on fabrication rates. [Alkaissi & McFarlane \(2023\)](#) reported that ChatGPT generates plausible-looking but entirely fictitious references. [Agrawal et al. \(2024\)](#) studied hallucinated citations across domains and models. More recently, [Xu et al. \(2026\)](#) benchmarked 13 LLMs on citation generation, finding hallucination

rates between 14% and 95% across vendors with no consistent relationship between batch size and hallucination rate. [Sakai et al. \(2026\)](#) identified nearly 300 published papers containing at least one hallucinated citation—most published in 2025—and proposed an OCR-based detection method. [Ansari \(2026\)](#) analyzed 100 hallucinated citations across 53 NeurIPS 2025 papers, finding that 66% were total fabrications and 100% exhibited compound failure modes where multiple fields were simultaneously incorrect. Their taxonomy distinguishes total fabrication, partial attribute corruption, identifier hijacking, placeholder hallucination, and semantic hallucination.

A.2 Field-level accuracy of LLM-generated bibliographic records

Studies of field-level accuracy differ along two axes: how citations are elicited and how errors are evaluated. Three data collection approaches have emerged: (a) open-ended generation, where an LLM is prompted to produce reference lists from scratch; (b) known-item retrieval, where the LLM is given an obfuscated description of a specific paper and asked to recover the full citation; and (c) verification of existing or synthetic citations against authoritative databases. Evaluation methods diverge correspondingly—per-field error rates, weighted composite scores, categorical classification, and binary existence checks—making cross-study comparison difficult.

Three studies adopt the open-ended generation approach. [Walters & Wilder \(2023\)](#) conducted the first systematic field-level study, prompting GPT-3.5 and GPT-4 to generate literature reviews across 42 multidisciplinary topics, then verifying the resulting 636 citations against Google Scholar, PubMed, Scopus, WorldCat, and other databases. Fabrication rates were 55% for GPT-3.5 and 18% for GPT-4. Among non-fabricated citations, numerical fields were most error-prone: volume/issue/pages were incorrect 34% of the time for GPT-3.5 and 13% for GPT-4, while dates showed 22% and 16% error rates respectively. Title accuracy was highest (94–97% correct). [Chelli et al. \(2024\)](#) prompted GPT-3.5, GPT-4, and Bard to generate references matching inclusion criteria for 11 systematic reviews in orthopedic surgery and related fields, using PubMed as ground truth. They defined a reference as hallucinated if two or more of three core fields—title, first author, year—were incorrect, a threshold that conflates partial corruption with fabrication. Title and journal accuracy exceeded 96% for non-hallucinated papers, but DOI accuracy was 16–20%. [Aljamaan et al. \(2024\)](#) developed a weighted Reference Hallucination Score (RHS) across 7 fields (4 major, 3 minor), with weights determined via Delphi consensus among librarians and physicians. They tested 6 systems including retrieval-augmented tools (Elicit, SciSpace); standalone LLMs scored at the maximum hallucination level, while retrieval-augmented tools scored near zero. The variation across these three studies is instructive: Walters uses post-hoc multi-database verification with binary fabricated/real judgments; Chelli thresholds on 2-of-3 core fields against PubMed; Aljamaan applies a Delphi-weighted composite score. Each captures different aspects of the same underlying problem.

[Szeider \(2026\)](#) adopted a different elicitation strategy: known-item retrieval with obfuscated citations at varying difficulty levels. Rather than asking the model to generate references, they provided descriptions of 104 known DBLP papers and tested whether Claude Sonnet 4.5 could recover correct BibTeX entries under three configurations—web search only, DBLP-mediated, and DBLP-unmediated. Evaluation used a 6-way categorical scheme (perfect match, wrong paper, not found, incomplete metadata, incomplete author, corrupted metadata) rather than per-field error rates. Unmediated database export achieved 82.7% perfect match versus 28.2% for the web baseline. The mediated approach eliminated metadata corruption but still dropped DOI, volume, and pages 36.5% of the time because the LLM reconstructed entries from memory rather than copying verbatim. This is the only study testing architectural design choices rather than comparing models.

A third group of studies evaluates citations through verification rather than generation-then-checking. [Algaba et al. \(2024\)](#) checked binary existence via Semantic Scholar fuzzy matching, finding that 12–36% of references generated by GPT-4, GPT-4o, and Claude 3.5 Sonnet did not correspond to any indexed paper. Their primary contribution is distributional: generated references had a median citation count 1,326 higher than actual references, quantifying how LLMs amplify the Matthew effect. [Chern et al. \(2023\)](#) verified only three

fields—title, authors, and year—via Google Scholar lookup as one component of a broader factuality framework, reporting 82.3% of LLM-generated citations as non-factual and that tool-augmented verification with GPT-4 achieved 95.24 F1 at detection. [Yuan et al. \(2026\)](#) introduced CiteAudit, a controlled perturbation benchmark with 9,442 citation instances where hallucinations are generated through targeted edits (keyword substitution, author perturbation, venue/year changes). Their multi-agent verification system achieved 0.903 F1 on real-world hallucinated citations, while general-purpose LLMs degraded to 0.331 F1 or below. These three studies differ fundamentally from the first group: they measure existence, factuality, or detection accuracy rather than per-field error rates across a comprehensive field set.

Across these studies, title is consistently the most accurate field when the cited paper exists, and DOI is consistently the worst—often fabricated or omitted entirely. Numerical metadata (volume, pages, issue) are error-prone across all models. Two recent studies extend these findings along complementary axes. [Niimi \(2025\)](#) generated 100 bibliographic records across 20 computer science domains with GPT-4.1 and found a strong log-linear correlation ($r = 0.75$) between citation count and metadata accuracy: papers with fewer than ~ 100 citations were predominantly hallucinated, while those exceeding $\sim 1,000$ citations were recalled near-verbatim, suggesting a memorization threshold tied to training-data redundancy. [Boudourides \(2026\)](#) introduced the concept of *structural hallucination*—systematic distortion of relational architecture invisible to sentence-level metrics—and evaluated gpt-4.1-mini on 50 COVID-19 publications against Dimensions.ai metadata. DOI reconstruction achieved F1 = 0.01, and 91.9% of ground-truth citations were omitted entirely, illustrating that without retrieval tools, LLMs fail to reconstruct bibliographic structure.

Methodological fragmentation further limits cross-study comparison. There is no agreed-upon minimal field set: studies evaluate 3 to 7+ fields with different inclusion criteria. Hallucination definitions range from binary fabricated/real ([Walters & Wilder, 2023](#)) to 2-of-3-field thresholds ([Chelli et al., 2024](#)) to weighted composite scores ([Aljamaan et al., 2024](#)) to 6-way categorical schemes ([Szeider, 2026](#)). Ground truth sources vary from post-hoc database search to pre-defined authoritative records to Delphi expert consensus. No standardized benchmark existed before CiteAudit ([Yuan et al., 2026](#)), and even CiteAudit checks only six fields, omitting volume and pages. Reported accuracy numbers are not directly comparable without normalizing for these differences.

A.3 Mitigation approaches

Retrieval-augmented generation ([Lewis et al., 2021](#)) and tool-augmented generation ([Nakano et al., 2022](#)) offer general strategies for grounding LLM outputs in external knowledge. Several systems have adapted these strategies specifically for citation accuracy.

On the verification side, the FacTool and CiteAudit systems discussed in Section A.2 double as mitigation tools; CiteAudit’s architecture is particularly relevant here because it cascades through memory lookup, web search, and a dedicated scholar agent, providing resilience when any single source fails. CheckIfExist ([Abbonato, 2026](#)) validates bibliographic entries against CrossRef, Semantic Scholar, and OpenAlex using a multi-stage matching algorithm with cascading fallback.

On the generation side, BibAgent ([Li et al., 2026](#)) constructs a benchmark of miscitation cases and uses retrieval-augmented “evidence committees” to verify claims against source documents. CiteLLM ([Hong et al., 2026](#)) generates context-aware queries and retrieves candidate papers with full-text semantic analysis, achieving 87.5% precision. CiteGuard ([Choi et al., 2026](#)) uses retrieval-augmented validation with full-text snippet retrieval to ground citation attribution. As shown in Section A.2, [Szeider](#)’s unmediated export approach eliminates metadata corruption entirely by bypassing the LLM during BibTeX construction.

All of these approaches bypass LLM parametric memory for bibliographic record construction. Grounding in authoritative databases, whether through retrieval augmentation, tool use, or direct export, consistently reduces both fabrication and field-level errors. Our work builds on this insight with `clibib`, which fetches authoritative BibTeX entries from the Zotero Translation Server to detect and correct hallucinated citations.

B LLM-as-judge rubric and borderline examples

Stage 2 of the evaluation pipeline (Section 3.6) uses an LLM judge to classify the 1,472 fields that Stage 1 could not resolve deterministically. The judge evaluates each field against two binary criteria via the autorubric framework (Rao & Callison-Burch, 2026), using Gemini 3.1 Flash Lite with structured output constrained to MET/UNMET/CANNOT_ASSESS per criterion. This section documents the full criteria definitions, verdict mapping, and borderline cases identified during human validation.

B.1 Criteria definitions

Criterion 1: partial_match. “Does the generated value partially match any ground-truth version?” The criterion is MET if there is meaningful overlap between the generated value and at least one ground-truth value. Examples of partial matches include: a truncated or reordered author list that shares most authors with the ground truth; an abbreviated or slightly different venue name referring to the same venue (e.g., “Proc. NeurIPS” vs. “Advances in Neural Information Processing Systems”); a title with minor wording differences, a missing subtitle, or formatting artifacts; a page range that overlaps or is a subset of the ground truth; a year off by one due to publication lag; or minor formatting differences in a DOI (e.g., URL prefix vs. bare identifier). The criterion is UNMET if the generated value has no meaningful overlap with any ground-truth version—the values refer to entirely different content or are fabricated with no recognizable connection to the ground truth.

Criterion 2: different_paper. “Does the generated value appear to come from a completely different paper than the ground truth?” The criterion is MET if the value clearly refers to a different work. Indicators include: author names sharing no overlap with any ground-truth version; a title describing a different topic or method; a DOI resolving to a different paper; a venue/year combination inconsistent with all ground-truth versions (e.g., “ICML 2020” vs. “NeurIPS 2017”); or volume/number/pages from a different journal issue. The criterion is UNMET if the value could plausibly refer to the same paper, even if imprecise or partially wrong. When uncertain, the conservative default is UNMET—a value is flagged as different-paper only when there is clear evidence of substitution.

B.2 Verdict mapping

The two criterion verdicts are combined into a single field label (P/S/F) according to Table 5. The priority logic is: (1) if partial_match is MET, the label is always P; (2) otherwise, if different_paper is MET, the label is S; (3) otherwise, the label is F. CANNOT_ASSESS is treated conservatively as UNMET, except that evidence of substitution (different_paper MET) overrides uncertainty about partial match.

Table 5: Verdict mapping from binary criteria to field labels.

partial_match	different_paper	Label
MET	any	P
UNMET	MET	S
UNMET	UNMET	F
UNMET	CANNOT_ASSESS	F
CANNOT_ASSESS	MET	S
CANNOT_ASSESS	UNMET	F
CANNOT_ASSESS	CANNOT_ASSESS	F

B.3 Few-shot examples in the judge prompt

The judge prompt includes three few-shot examples to calibrate the criteria:

1. **Partial match (author).** Generated: “Vaswani, A. et al.”; ground truth: “Vaswani, Ashish and Shazeer, Noam and Parmar, Niki ... and Polosukhin, Illia.” The

first author overlaps but co-authors are truncated. Verdict: `partial_match MET`, `different_paper UNMET` → P.

2. **Substitution (venue).** Generated: “ICML 2020”; ground truth: “Advances in Neural Information Processing Systems (NeurIPS) 2017.” Different conference and different year indicate a value from a different paper. Verdict: `partial_match UNMET`, `different_paper MET` → S.
3. **Fabrication (DOI).** Generated: “10.1234/fake.5678”; ground truth: “10.1038/s41586-020-1234-5.” No structural overlap; the registrant code “1234” does not exist. The value appears invented rather than drawn from another real paper. Verdict: `partial_match UNMET`, `different_paper UNMET` → F.

B.4 Borderline cases from human validation

Human validation of 521 Stage 2 fields (Section 6.4) identified two systematic disagreement patterns between the LLM judge and the human annotator. These account for 66 of 102 total disagreements (64.7%).

Entry-type mismatches (F vs. P). The LLM judge labeled all entry-type mismatches as fabrications (F), treating any type difference as invented content. The human annotator labeled them as partial matches (P), reasoning that semantically related BibTeX types for the same paper—e.g., `misc` vs. `article`, or `misc` vs. `inproceedings`—reflect an imprecise but not fabricated classification. This pattern accounts for 37 of the 44 F→P disagreements. For example, a generated `misc` entry for a paper whose ground truth is `article` was labeled F by the judge (no textual overlap between the type strings) but P by the human (the paper exists, and `misc` is a common catch-all type in BibTeX).

Preprint-vs-published DOI versioning (S vs. P). When a generated DOI points to an arXiv preprint (e.g., 10.48550/arXiv.2510.16227) and the ground truth contains the published version’s DOI (e.g., 10.1162/TACL.a.611), the LLM judge labeled the mismatch as a substitution (S)—the structurally different registrant prefixes led it to conclude the value came from a different paper. The human annotator labeled these as partial matches (P), recognizing that a preprint DOI and its corresponding published DOI refer to the same underlying work. This pattern also extends to other preprint servers: e.g., a ChemRxiv DOI (10.26434/chemrxiv-2025-90r12) vs. its published journal DOI (10.1021/acs.chemmater.5c02918). Conversely, a small number of cases show the opposite error: DOIs with similar prefixes (e.g., two 10.1002/adma.* identifiers) were labeled P by the judge but S by the human, because the suffixes resolved to different papers despite the shared registrant.

C Failure mode examples

Table 6 shows one representative entry for each of the two error modes identified in Section 4.6.

The wholesale-substitution entry illustrates this mode clearly: GPT-5 returned metadata for a real paper by the same first author (Yamashita), in the same disease area (upper urinary tract urothelial carcinoma), published the same year (2016)—but in the wrong journal. The generated volume, number, and pages are internally consistent with the substituted journal rather than the ground truth, indicating that the model retrieved a coherent but incorrect record.

The isolated-error entry illustrates the opposite pattern: Gemini correctly identified a well-known NeurIPS 2012 paper, matched all identity fields, but fabricated a page range offset by exactly 9 pages (the correct page span). This near-miss pattern, where the page count is correct but the starting page is wrong, is characteristic of parametric memory that encodes document length but not exact position within proceedings.

Table 6: Representative examples of the two failure modes. **Wholesale substitution:** GPT-5 returns a coherent but wrong paper by the same first author, same year, same disease—a plausible near-miss. **Isolated field error:** Gemini 3 Flash identifies the correct paper but fabricates a page range offset by exactly 9 pages.

Mode	Field	Generated	Ground truth	Label
<i>Wholesale substitution</i> (GPT-5, medicine, low-citation)				
	title	Clinical implications of intravesical recurrence after radical nephroureterectomy...	Impact of relapse site on oncological outcomes after radical nephroureterectomy...	S
	venue	International Journal of Urology	Journal of Clinical Oncology	S
	volume	23	34	F
	number	5	2.suppl	F
	pages	378–384	426–426	F
	doi	10.1111/iju.13054	10.1200/jco.2016.34.2.suppl.426	S
<i>Isolated field error</i> (Gemini 3 Flash, AI, popular)				
	title	Learning to Discover Social Circles in Ego Networks	Learning to Discover Social Circles in Ego Networks	C
	author	Julian McAuley, Jure Leskovec	Julian J. McAuley, Jure Leskovec	C
	year	2012	2012	C
	venue	Advances in Neural Information Processing Systems 25	Advances in Neural Information Processing Systems 25	C
	pages	539–547	548–556	F

D Two-stage revision prompt

Listing 1 reproduces the revision prompt used in Phase 2 of the two-stage integration (Algorithm 1, Section 5.2). The same prompt is used for all three models. The placeholders {generated_bibtex} and {clibib_bibtex} are filled with the model’s baseline output and the clibib lookup result, respectively.

Listing 1: Two-stage revision prompt template.

```

You previously generated a BibTeX entry via web search. An authoritative database
lookup (querying DBLP and CrossRef) returned a verified entry. Compare the
titles of both entries carefully. If the database entry is for the SAME paper (
titles match modulo latex markups), replace your values with the database
values for: authors, title, year, venue/booktitle/journal, volume, number,
pages, and DOI. Only keep your original value for a field if the database entry
is missing that field entirely. If the database entry appears to be for a
DIFFERENT paper (titles do not match), keep your original entry unchanged.
Return the final BibTeX inside <bibtex></bibtex> tags.

Your generated BibTeX:
{generated_bibtex}

BibTeX from database lookup:
{clibib_bibtex}

```

E Dataset construction details

This appendix collects supporting tables, figures, and analyses from the dataset construction pipeline (Section 3.4).

E.1 Venue listing

Table 7 lists the 10 venues selected for each of the four domains. Venues marked with † require topic filtering to restrict to the relevant subfield.

Table 7: Venues per domain. Venues marked with † require topic filtering (quantum computing or materials chemistry subsets of broader journals).

Domain	Venues
AI	NeurIPS, ICML, ICLR, ACL, EMNLP, CVPR, AAAI, NAACL, JMLR, TACL
Medicine	NEJM, The Lancet, JAMA, BMJ, Nature Medicine, Ann. Internal Med., PLOS Medicine, Lancet Digital Health, J. Clinical Oncology, Circulation
Materials Science	Nature Materials, Adv. Materials, ACS Nano, Nano Letters, Chem. Materials, Acta Materialia, Adv. Energy Materials, Materials Today, JACS [†] , npj Comp. Mat.
Quantum Comp.	PRL [†] , Nature Physics, PRX Quantum, Quantum, npj Quantum Info., Phys. Rev. A, Quantum Sci. Tech., Nature Comms. [†] , IEEE Trans. QE, New J. Physics [†]

E.2 Dataset composition and citation distribution

Figure 3 shows the final dataset composition after quality filtering. Each domain initially targeted 100 popular, 100 low-citation, and 50 recent papers; final counts reflect removal of 66 non-research items. Figure 4 confirms that the three citation tiers are well-separated, spanning five orders of magnitude in cited-by counts.

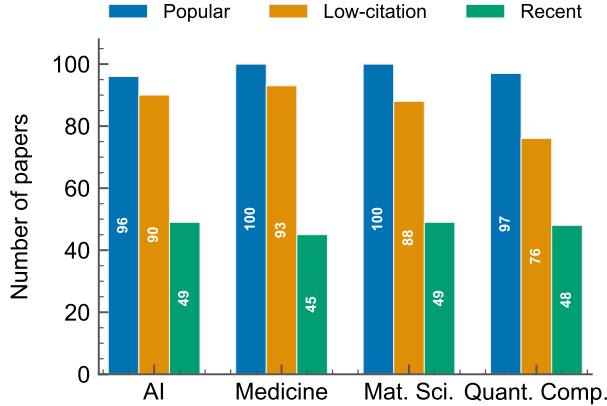


Figure 3: Paper counts by domain and citation tier after quality filtering. Each domain initially targeted 100 popular, 100 low-citation, and 50 recent papers; final counts reflect removal of 66 non-research items during the data quality audit.

E.3 Post-hoc 2025 overlap analysis

Because knowledge cutoffs vary across models (Table 1), some AI-domain 2025 papers fall within Claude Sonnet 4.6’s knowledge window (August 2025 cutoff) but remain post-cutoff for GPT-5 (October 2024) and Gemini 3 Flash (January 2025). A post-hoc check suggests this overlap does not confound results: field-level accuracy drops from 2025 to 2026 AI papers for all three models, including GPT-5 (59.7% → 34.6%), whose October 2024 cutoff predates all papers in this tier. The uniform drop indicates that the 2025–2026 gap reflects search indexing recency rather than parametric memory.

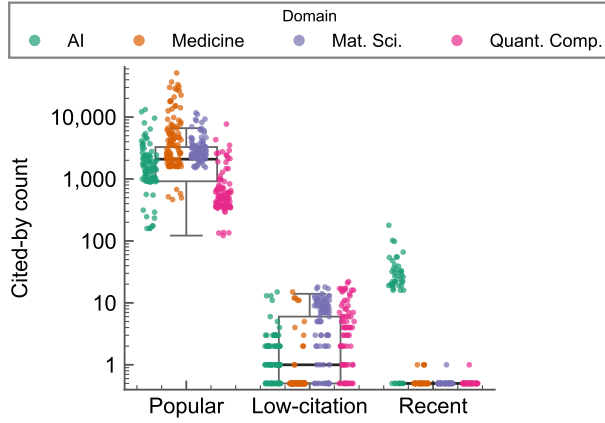


Figure 4: Distribution of cited-by counts across citation tiers (log scale). Popular papers have median citations of 500–3,567; low-citation papers cluster near 0–7; recent papers are predominantly uncited. The clear separation confirms that the tiers probe distinct regions of model familiarity.

E.4 Version landscape

Table 8 and Figure 5 summarize the version composition of the 997 initially sampled papers. The version composition varies substantially across domains: AI is proceedings-dominated (86.2%), Quantum Computing has the richest multi-version landscape (44.8% with both journal and arXiv versions), while Medicine and Materials Science are almost entirely journal-only.

Table 8: Version landscape of the 997 papers. Columns report the number of papers with at least one version of the given type. Multi-ver. counts papers with more than one fetchable version (i.e., both a published venue version and an arXiv preprint). Total ver. is the sum of all fetchable versions across papers.

Domain	<i>n</i>	Proc.	Journal	arXiv	Multi-ver.	Total ver.
AI	247	218	29	5	5 (2.0%)	252
Medicine	250	0	250	1	1 (0.4%)	251
MatSci	250	0	250	16	16 (6.4%)	266
QuantComp	250	0	250	112	112 (44.8%)	362
Total	997	218	779	134	134 (13.4%)	1,131

E.5 BibTeX retrieval outcomes

Figure 6 shows BibTeX retrieval outcomes by domain. Quantum Computing and Materials Science achieve >95% success rates. Medicine’s lower rate reflects paywalled journal articles that the Zotero Translation Server cannot resolve, with many title-mismatch failures on fallback queries.

E.6 Cross-validation details

We cross-validate clibib results against domain-specific authoritative databases. For AI and Quantum Computing papers (497 total), we query the DBLP API with the paper title, accept matches at ≥ 0.85 title token overlap, and compare venue, year, pages, and author fields against the clibib result, flagging discrepancies. The DBLP query matched 216 of 247 AI papers (87.4%) but only 22 of 250 Quantum Computing papers (8.8%), because DBLP focuses on computer science venues rather than physics journals. For Medicine papers (250

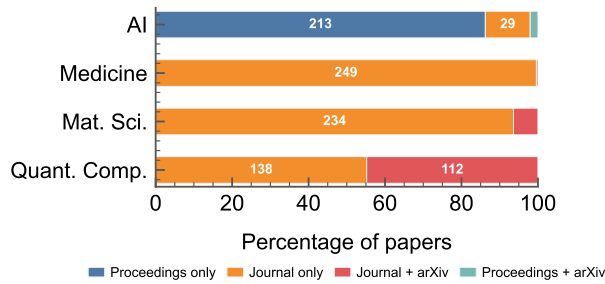


Figure 5: Version composition of papers by domain. Each bar shows the proportion of papers in each version combination. Quantum Computing has the highest multi-version rate (44.8%), driven by the physics preprint culture. AI is proceedings-dominated (86.2%), while Medicine and Materials Science are almost entirely journal-only.

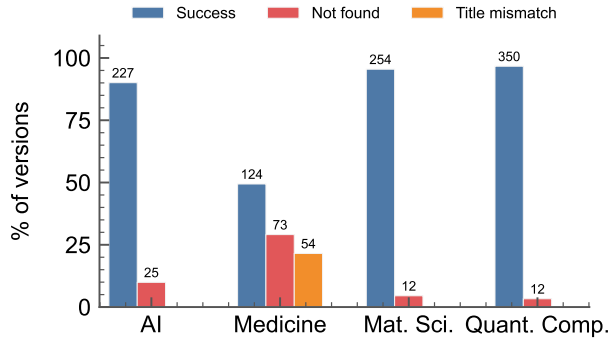


Figure 6: BibTeX retrieval outcomes by domain. Quantum Computing and Materials Science achieve >95% success. Medicine’s lower rate reflects paywalled journal articles that Zotero cannot resolve, with many title-mismatch failures on fallback queries.

total), we query PubMed via the NCBI E-utilities API using the DOI and compare journal, year, volume, pages, and author fields; PubMed matched 148 papers (59.2%). Materials Science has no single domain-specific authoritative database and relies on `clibib` results alone. Each cross-validation record stores the match status, parsed fields, and field-level discrepancies.

E.7 Canonical field coverage

Figure 7 shows the canonical field coverage rates by domain. Core fields (DOI, title, author, year) exceed 80% in all domains except Medicine. Booktitle coverage is concentrated in AI (conference proceedings); journal dominates the other three domains.

E.8 Venue synonym examples

The dominant patterns among the 94 genuinely incorrect venue cases are preprint-vs-published substitutions (“arXiv preprint arXiv:1802.03268” for a paper published at ICML), confusable sibling journals within the same publisher family (“JCO Oncology Practice” vs. “Journal of Clinical Oncology”), and cross-venue substitutions within the same subfield (“Physical Review A” vs. “Physical Review Letters”). These cases reflect model errors rather than normalization gaps: the synonym table resolves formatting variation effectively, while the Stage 2 LLM judge handles the remaining semantic distinctions.

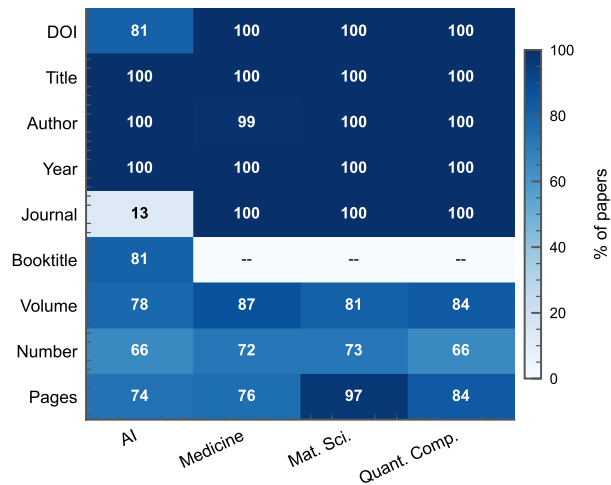


Figure 7: Canonical field coverage by domain (percentage of papers with a resolved value). Core fields (DOI, title, author, year) exceed 80% in all domains except Medicine. Booktitle is concentrated in AI (conference proceedings); journal dominates the other three domains. Medicine has the lowest coverage due to its lower BibTeX retrieval rate.

F Supplementary evaluation results

This appendix collects figures, tables, and extended discussion moved from Section 4 to conserve space in the main body.

F.1 Per-field accuracy by model

Figure 8 shows per-field accuracy broken out by model. Gemini 3 Flash leads on all nine fields; GPT-5 and Claude are competitive on author and pages. The per-field gap is largest for entry type (96.9% vs. 90.3% GPT-5), venue (94.7% vs. 84.8%), and volume (92.6% vs. 82.4%).

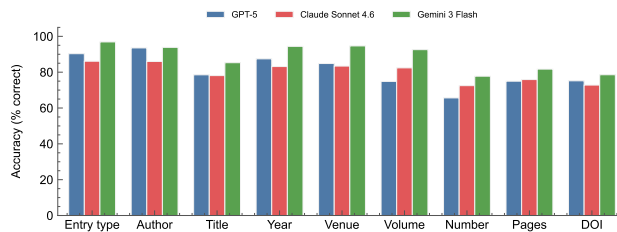


Figure 8: Per-field accuracy by model. Gemini 3 Flash leads on all nine fields; GPT-5 and Claude are competitive on author and pages.

F.2 Domain accuracy

Figure 9 shows per-domain accuracy by model. Medicine is the hardest domain (77.7%), while Materials Science and Quantum Computing are easiest (86.8% each). Gemini 3 Flash leads in all four domains, with the largest margin in medicine (13.3 pp over GPT-5).

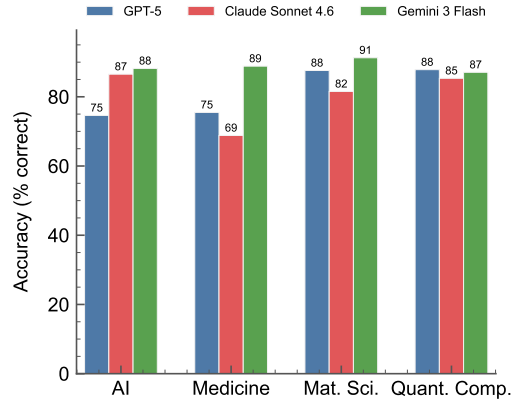


Figure 9: Per-domain accuracy by model. Medicine is the hardest domain; materials science and quantum computing are easiest.

F.3 Search accuracy by invocation

Figure 10 shows accuracy conditioned on whether the model invoked its search tool. GPT-5 achieves 85.1% without search on 49 entries (predominantly popular papers), consistent with strong parametric memory. Gemini and Claude without search produce all-missing fields; these entries are retained in all analyses.

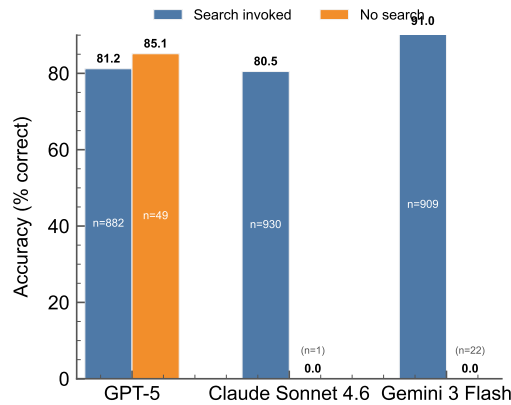


Figure 10: Accuracy conditioned on search invocation, by model. GPT-5 achieves 85.1% without search on 49 entries (predominantly popular papers). Gemini and Claude without search produce all-missing fields (0% accuracy); these 22 and 1 entries, respectively, are included in all analyses.

F.4 Per-field accuracy by citation tier

Table 9 breaks out the tier effect by field. Pages degrades most (−51.8 pp from popular to recent), followed by DOI (−37.8 pp) and volume (−32.6 pp). Number shows the smallest drop (−14.8 pp), though its baseline accuracy is already low (74.1%). Author is the most robust to recency (−17.5 pp), likely because author names are prominent in search results.

F.5 Human agreement details

Table 10 reports per-field inter-annotator agreement between the LLM judge and the human annotator on 521 Stage 2 fields. Year achieves perfect agreement ($\kappa = 1.0$); entry_type

Table 9: Per-field accuracy (% C) by citation tier, pooled across models. 95% bootstrap CIs in brackets. Δ is the popular-to-recent drop in percentage points.

Field	Popular	Low-cit.	Recent	Δ
Entry type	99.4 [99.0, 99.7]	92.0 [90.2, 93.9]	72.4 [68.4, 76.3]	-27.0
Author	97.3 [95.8, 98.6]	90.4 [88.0, 92.6]	79.8 [75.6, 83.8]	-17.5
Title	90.6 [88.0, 93.0]	74.4 [70.4, 78.2]	71.6 [66.5, 76.4]	-19.0
Year	98.7 [97.9, 99.4]	88.3 [85.8, 90.6]	67.0 [62.5, 71.6]	-31.7
Venue	97.4 [96.0, 98.6]	88.0 [85.6, 90.4]	67.0 [62.3, 72.1]	-30.4
Volume	89.8 [87.1, 92.2]	82.5 [79.8, 85.3]	57.2 [50.6, 63.4]	-32.6
Number	74.1 [70.6, 77.5]	71.5 [68.2, 74.9]	59.3 [51.9, 67.4]	-14.8
Pages	93.2 [90.7, 95.5]	74.9 [71.1, 78.9]	41.4 [34.9, 48.1]	-51.8
DOI	90.0 [87.4, 92.4]	73.0 [69.1, 76.7]	52.2 [47.1, 57.4]	-37.8
Overall	92.7	82.0	65.0	-27.7

shows zero agreement due to the systematic pattern where the LLM judge labels all type mismatches as fabrications while the human labels them as partial matches. Excluding entry_type and DOI raises κ to 0.84.

Table 10: Per-field inter-annotator agreement between the LLM judge and human annotator on 521 Stage 2 fields. Fields are ordered by descending κ . † marks fields with systematic disagreement patterns (see text).

Field	n	Agreement (%)	κ
year	56	100.0	1.000
number	27	92.6	0.816
author	37	89.2	0.798
pages	65	81.5	0.674
title	157	96.8	0.536
volume	29	79.3	0.461
venue	35	80.0	0.426
doi [†]	76	61.8	0.217
entry_type [†]	39	5.1	0.000
All fields	521	80.4	0.671
Excl. entry_type	482	86.5	0.772
Excl. entry_type & doi	406	91.1	0.838

F.6 Error-type distribution per field

Figure 11 shows the full error-type distribution per field across all models. Missing values dominate for number (24.9%) and DOI (16.0%). Fabrication is highest for pages (4.5%) and entry type (3.7%). Partial matches concentrate in title (13.9%), where minor formatting differences account for most non-correct labels.

F.7 Error-type composition by model

Figure 12 shows the error composition by model, excluding correct and not-applicable fields. Missing is the dominant error type for all three models. Claude has the highest missing rate (13.8%); GPT-5 has the highest fabrication rate (2.2%); Gemini has the lowest error rates across all categories.

F.8 Author matching sensitivity

The baseline evaluation matches authors by first-author last name only (Section 3.6), a deliberately lenient criterion congruent with the elicitation prompt, which provides the title and first author. Table 11 reports accuracy under stricter criteria. Requiring all authors'

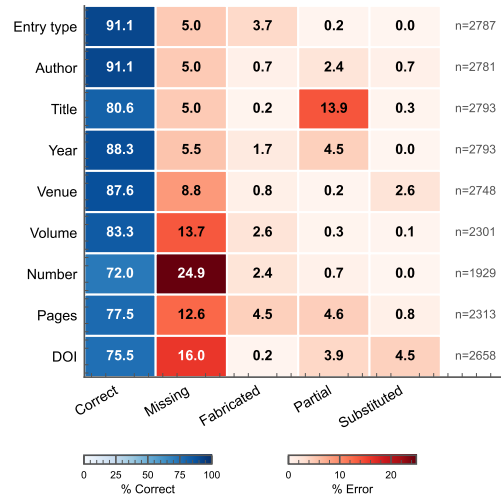


Figure 11: Error-type distribution per field (all models combined). Each row shows the fraction of evaluable fields receiving each label. Author, entry type, and year have the highest correct rates; number and DOI are most error-prone.

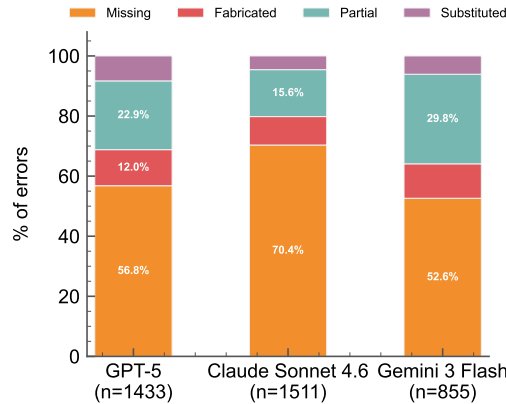


Figure 12: Error composition by model (excluding correct and not-applicable fields). Missing is the dominant error type for all models. Claude has the highest missing rate; GPT-5 has the highest fabrication rate.

last names to match as a set reduces accuracy from 90.7% to 79.6% (−11.1 pp), with consistent drops across models (9.0–13.4 pp) and tiers: popular 97.0%→87.4%, low-citation 89.6%→77.0%, recent 79.8%→68.2%. Ordered matching barely differs from set matching (79.5% vs. 79.6%), confirming that models preserve author order when they recover the correct names.

Among the 310 entries that lose their correct label under set matching, truncation is the dominant failure mode (33%): the model finds the right paper but omits co-authors or emits an abbreviated list. A further 27% reflect name-normalization artifacts—formatting differences in multi-word surnames, LaTeX accent encoding, or prefix placement (e.g., “Ó Séaghdha” vs. “Séaghdha, Diarmuid Ó”)—rather than genuinely incorrect authors. The remaining 40% are split between entries with extra authors not in the ground truth (17%) and true substitutions where the last names have low overlap (23%). These results confirm that first-author matching inflates reported author accuracy; the gap is driven primarily by co-author completeness rather than retrieval of the wrong paper.

Table 11: Author-field accuracy under increasingly strict matching criteria. The baseline evaluation matches only the first author’s last name; stricter criteria require all authors’ last names to match. The 11.1 pp drop from first-author to set matching reflects truncated author lists and name-normalization artifacts rather than wholesale substitution (see text). Brackets denote 95% bootstrap confidence intervals.

Matching criterion	GPT-5	Claude Sonnet 4.6	Gemini 3 Flash	All
First-author last name	93.1 [91.3, 94.6]	85.6 [83.4, 87.8]	93.4 [91.7, 95.1]	90.7 [89.6, 91.7]
All last names (set)	82.3 [79.8, 84.6]	76.6 [73.9, 79.2]	80.0 [77.4, 82.6]	79.6 [78.1, 81.1]
All last names (ordered)	82.3 [79.8, 84.6]	76.4 [73.6, 79.1]	79.9 [77.3, 82.5]	79.5 [78.0, 81.0]
Δ (first-author $\rightarrow$ set)	−10.8	−9.0	−13.4	−11.1

F.9 Accuracy by citation tier and model

Table 12 provides accuracy by citation tier and model with 95% bootstrap confidence intervals. Popular papers achieve 92.7% field-level accuracy; low-citation papers drop to 82.0%; recent post-cutoff papers fall to 65.0%. Fully correct entry rates follow the same gradient: 60.3–68.2% for popular, 42.7–49.6% for low-citation, and 23.0–39.3% for recent.

Table 12: Accuracy by citation tier and model. Field-level: % of evaluable fields labeled C. Fully correct: % of entries with all evaluable fields correct. 95% bootstrap CIs in brackets (2,000 cluster-resampled iterations).

Tier	Model	Field-level	Fully correct
Popular	GPT-5	91.7 [90.3, 92.9]	60.3 [55.7, 65.4]
	Claude 4.6	93.4 [92.1, 94.6]	68.2 [63.6, 72.5]
	Gemini 3 Fl.	93.1 [91.6, 94.4]	64.1 [59.3, 68.7]
	<i>Pooled</i>	92.7 [91.6, 93.7]	—
Low-citation	GPT-5	81.4 [78.6, 84.1]	42.7 [37.2, 47.8]
	Claude 4.6	77.8 [74.3, 81.3]	49.0 [43.8, 54.2]
	Gemini 3 Fl.	86.8 [84.2, 89.1]	49.6 [44.4, 55.0]
	<i>Pooled</i>	82.0 [79.7, 84.1]	—
Recent	GPT-5	56.9 [51.7, 62.1]	23.0 [17.3, 29.3]
	Claude 4.6	54.7 [48.4, 60.9]	28.8 [22.5, 35.1]
	Gemini 3 Fl.	83.4 [80.2, 86.2]	39.3 [32.5, 46.6]
	<i>Pooled</i>	65.0 [61.5, 68.7]	—
All	<i>Pooled</i>	83.6	—

F.10 Search depth details

The number of sources consulted per entry differs by an order of magnitude. GPT-5 retrieves a median of 50 sources per entry (mean 68.2, range 2–372), reflecting its iterative tool-call architecture that issues multiple web queries per prompt. Claude Sonnet 4.6 consults a median of 10 sources (mean 13.0), and Gemini 3 Flash retrieves a median of 8 sources (mean 7.6). For all three models, source counts increase for harder tiers: GPT-5’s median rises from 30 (popular) to 64–66 (low-citation/recent); Claude’s median rises from 10 to 13. These differences reflect each vendor’s search implementation rather than a configuration choice.

Because vendor APIs expose search as an opaque tool, we cannot determine whether BibTeX fields were transcribed from a retrieved page or reconstructed from parametric memory after search. The “last mile” failure pattern—correctly locating a paper but incorrectly reporting its metadata—is discussed in Section 6. Resolving grounded-versus-reconstructed attribution would require output-provenance tracing, which no current API supports.

G Supplementary mitigation results

This appendix collects figures, tables, and extended discussion moved from Section 5 to conserve space in the main body.

G.1 Single-stage algorithm

Algorithm 2 details the single-stage clibib integration protocol. The `clibib_lookup` function is registered alongside the native web search tool, and the model drives the tool loop: it decides which query to send, interprets the response, and produces the final BibTeX. The loop runs for up to 3 tool rounds; if no final response is produced, a forced completion call elicits a text response. Gemini is excluded from the single-stage condition because Google’s API does not allow GoogleSearch grounding and function declarations in the same request.

Algorithm 2 Single-stage clibib integration

Require: Paper description d , model $M \in \{\text{GPT-5}, \text{Claude}\}$
Ensure: BibTeX entry b

```

1:  $tools \leftarrow [\text{WEBSEARCH}, \text{CLIBIBLOOKUP}]$ 
2:  $prompt \leftarrow \text{RENDERPROMPT}(d)$  ▷ KR-nat + clibib instructions
3:  $messages \leftarrow [prompt]$ 
4: for  $round = 1$  to 3 do
5:    $response \leftarrow M.GENERATE(messages, tools)$ 
6:   if  $response$  contains text (no pending tool calls) then
7:     return  $\text{EXTRACTBIBTEX}(response)$ 
8:   end if
9:   for each tool call  $c$  in  $response$  do
10:    if  $c.name = \text{CLIBIBLOOKUP}$  then
11:       $r \leftarrow \text{CLIBIBLOOKUP}(c.query)$  ▷ Returns BibTeX or not_found
12:      Append  $(c, r)$  to  $messages$ 
13:    end if
14:  end for
15: end for
16:  $response \leftarrow M.GENERATE(messages, [\text{WEBSEARCH}])$  ▷ Force text
17: return  $\text{EXTRACTBIBTEX}(response)$ 
```

G.2 Single-stage model comparison

Figure 13 compares overall accuracy before and after single-stage clibib augmentation for GPT-5 and Claude. Both models improve: GPT-5 by +5.3 pp (81.4% to 86.7%) and Claude by +3.6 pp (80.4% to 84.0%).

G.3 Per-field before/after comparison

Figure 14 shows per-field accuracy changes with single-stage clibib augmentation. Number shows the largest gain (+19.1 pp), followed by volume (+9.0 pp), pages (+7.9 pp), and DOI (+7.2 pp). Identity fields (author, title, year, venue) show modest improvements, reflecting that these fields are already relatively accurate at baseline.

G.4 Correction rates by error type

Table 13 reports correction rates stratified by the original error label for the single-stage condition (GPT-5 and Claude). Missing fields had the highest correction rate (49.2%): clibib supplied metadata for fields the model omitted entirely. Fabricated and substituted fields were corrected at lower rates (18.1% and 14.9%), because these errors often involve wholesale entry substitution where the lookup query itself targets the wrong paper.

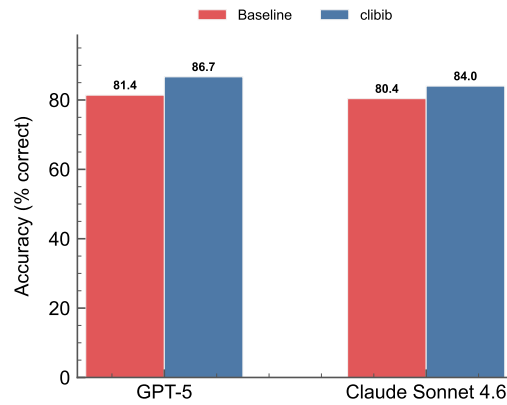


Figure 13: Overall accuracy (percentage of evaluable fields labeled C) before and after single-stage clibib augmentation (GPT-5 and Claude only). Both models improve: GPT-5 by +5.3 pp and Claude by +3.6 pp.

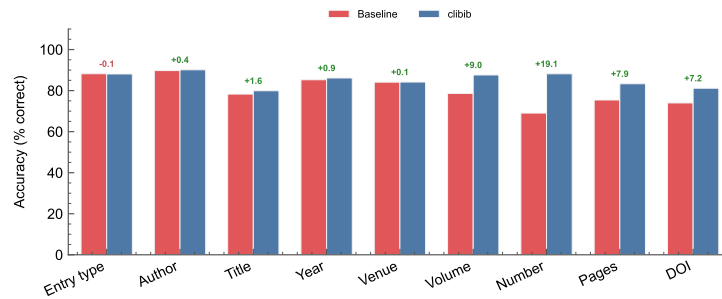


Figure 14: Per-field accuracy before (baseline) and after single-stage clibib-augmented generation (GPT-5 and Claude only). Annotations show the net change in percentage points. Number (+19.1 pp), volume (+9.0 pp), and pages (+7.9 pp) show the largest gains.

G.5 Query routing details

Internally, clibib routes queries through two Zotero Translation Server endpoints. URL queries (including normalized arXiv, alphaxiv, and HuggingFace links) are sent to the server’s /web endpoint, which scrapes the target page and extracts structured metadata. All other queries—DOIs, arXiv IDs, ISBNs, PMIDs, and free-text titles—are sent to the /search endpoint, which resolves identifiers against the server’s backend databases (CrossRef, DBLP, PubMed, and others). DOI-style URLs are an exception: the DOI is extracted from the URL and rerouted to /search for deterministic resolution rather than page scraping. When /search returns an empty response, clibib falls back to the CrossRef API (api.crossref.org/works), querying with the original text and retrieving up to 10 candidates. When either endpoint returns multiple candidates (an HTTP 300 response from Zotero, or multiple CrossRef hits), clibib ranks them by token-level Jaccard similarity between the query and each candidate’s title, with a substring-match tiebreaker, and selects the top-ranked candidate. If only one candidate is returned, it is accepted without ranking. In all cases, the resolved metadata is returned as Zotero JSON items, which clibib converts to BibTeX via the server’s /export endpoint. The output fields depend on what the upstream database deposited; clibib performs no field synthesis or inference beyond sanitizing the citation key to alphanumeric characters.

Table 13: Correction rates by original error type (GPT-5 and Claude only). Correction = baseline error that becomes C with single-stage clibib augmentation.

Error type	Baseline errors	Corrected	Rate (%)
Missing (M)	1,877	924	49.2
Partial (P)	564	268	47.5
Fabricated (F)	315	57	18.1
Substituted (S)	188	28	14.9
All errors	2,944	1,277	43.4

G.6 Correction by error type

Figure 15 shows correction rates by original error type for each model in the single-stage condition. GPT-5 shows higher correction rates than Claude across all error types, particularly for missing fields (53.7% vs. 42.6%).

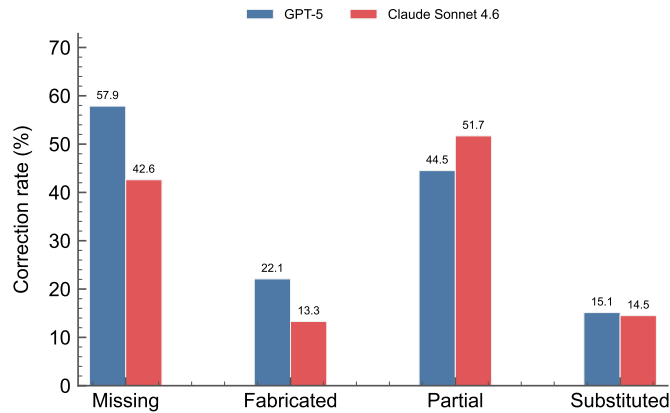


Figure 15: Correction rate by original error type, per model (single-stage clibib, GPT-5 and Claude only). Missing fields are corrected most often (49.2% overall). GPT-5 shows the highest correction rates for missing and partial errors.

G.7 Lookup coverage and failure modes

Table 14: clibib lookup coverage by model, domain, and citation tier (single-stage, GPT-5 and Claude only). “Lookups” counts individual clibib.lookup calls; “Success” is the fraction that returned a valid BibTeX entry; “Not found” is the fraction with no match in any backend database.

Condition	Lookups	Success (%)	Not found (%)
Overall	2,617	88.9	11.1
GPT-5	1,432	89.7	10.3
Claude Sonnet 4.6	1,185	87.9	12.1
AI	864	81.8	18.2
Medicine	642	87.5	12.5
Mat. Sci.	577	93.4	6.6
Quant. Comp.	534	97.2	2.8
Popular	1,071	89.4	10.6
Low-citation	935	90.8	9.2
Recent	611	85.3	14.7

Table 14 reports clibib lookup coverage across models, domains, and citation tiers. Of 2,617 individual `clibib_lookup` calls, 88.9% returned a valid BibTeX entry and 11.1% returned `not_found`. Coverage varies more across domains than across models: quantum computing lookups succeed at 97.2%, materials science at 93.4%, and medicine at 87.5%. AI has the lowest success rate (81.8%). The higher `not_found` rate in AI reflects the prevalence of workshop papers and preprints that lack DOIs or are not yet indexed in CrossRef or DBLP. Among tiers, recent papers show the highest `not_found` rate (14.7%) compared with popular (10.6%) and low-citation (9.2%), consistent with indexing lag for newly published work.

When `clibib` returns `not_found`, the model retains its original web-search-derived BibTeX unchanged—a no-op fallback that preserves baseline accuracy but forgoes the correction opportunity. At the entry level, 97.9% of entries had at least one successful lookup, meaning only 2.1% of entries received no authoritative data at all.

Residual errors after `clibib` augmentation fall into two categories: (1) the coverage gap just described, which limits `clibib`’s reach for recent and niche publications; and (2) substitution errors, where the model retrieved the wrong paper entirely (Section 4.6), which remain largely uncorrected as noted above. Regression rates nonetheless remain modest (Section 5.3).

G.8 Domain and tier effects

AI showed the largest domain-level gain (+8.7 pp, from 80.6% to 89.3%), followed by medicine (+5.6 pp, from 72.1% to 77.8%) and quantum computing (+2.8 pp, from 86.6% to 89.4%). Materials science showed a modest gain (+0.8 pp, from 84.6% to 85.4%).

Among tiers, low-citation papers showed the largest improvement (+7.4 pp, from 79.6% to 87.0%), followed by recent papers (+4.5 pp, from 55.8% to 60.4%) and popular papers (+1.8 pp, from 92.5% to 94.3%). The smaller gain for popular papers reflects their already-high baseline accuracy.

G.9 Correction patterns by model

Per-model analysis on GPT-5 and Claude—the two models with results under all three conditions—reveals distinct correction profiles (Figure 16). GPT-5 shows a consistent two-stage advantage across all four error types: substituted errors see the largest gain (+42.0 pp, from 15.1% to 57.1%), followed by fabricated (+22.1 pp), partial (+19.2 pp), and missing (+11.7 pp). Claude shows a two-stage advantage for fabricated errors (+32.2 pp) but a *reversal* for missing fields: single-stage corrects 42.6% of missing fields compared to 32.2% for two-stage ($\Delta = -10.4$ pp). The single-stage tool loop can proactively query `clibib` to fill gaps in the generated entry, while the two-stage revision prompt focuses on correcting existing fields rather than adding absent ones.

G.10 Strategy agreement and disagreement

We compare single-stage and two-stage labels on the same 15,402 evaluable fields for GPT-5 and Claude (Figure 17). The two strategies assign the same label in 88.3% of cases. When they disagree, two-stage wins $2.4\times$ more often than single-stage (7.6% vs. 3.1% of fields). Among the 482 fields where only single-stage is correct, 78.8% of the two-stage errors are missing-field labels, consistent with the revision prompt’s tendency to omit fields rather than fabricate them. Among the 1,084 fields where both strategies err, 86.3% share the same error type—shared failure modes from hard-to-retrieve papers rather than independent mistakes. The full label transition matrix (Figure 18) confirms that the dominant off-diagonal flow is between C and M in both directions: 630 fields transition from single-stage M to two-stage C (corrections); 380 transition from single-stage C to two-stage M (regressions to missing).

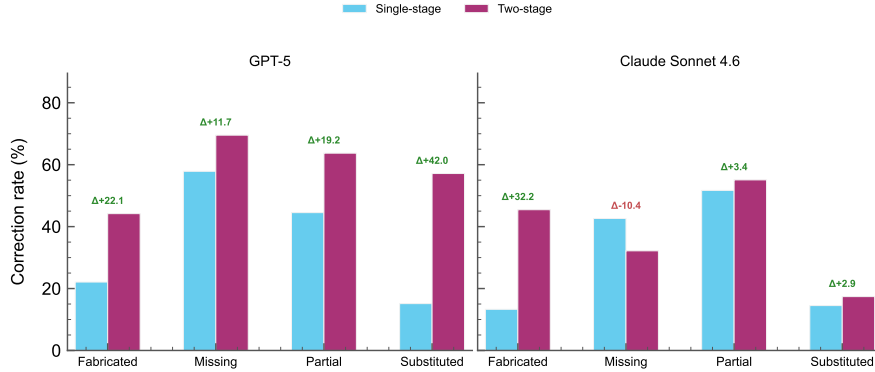


Figure 16: Per-error-type correction rates for single-stage vs. two-stage integration, by model. GPT-5 shows a consistent two-stage advantage across all error types; Claude shows a reversal for missing fields ($\Delta = -10.4$ pp), where single-stage corrects more effectively.

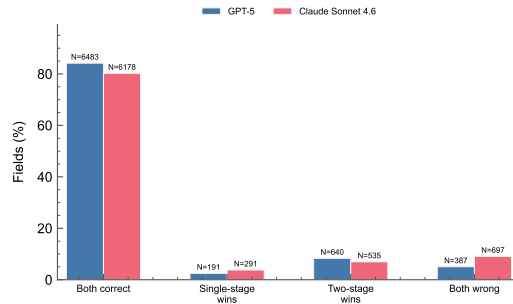


Figure 17: Concordance and discordance between single-stage and two-stage integration on 15,402 evaluable fields (GPT-5 and Claude). Two-stage wins $2.4\times$ more often than single-stage when the strategies disagree.

G.11 Two-stage ablation figures

Figures 19–21 provide supplementary visualizations for the two-stage ablation analysis. Figure 19 compares field-level accuracy across all three integration strategies. Figure 20 shows per-field accuracy before and after two-stage revision, with number (+22.1 pp), DOI (+14.9 pp), and pages (+13.0 pp) showing the largest gains. Figure 21 breaks down two-stage correction rates by error type and model.

H Extended comparison with prior work

This appendix provides detailed per-study comparisons that complement the summary in Section 6.3.

H.1 Field-level accuracy comparisons

Walters & Wilder (2023) reported 94–97% title accuracy; we observe 80.6%. Chelli et al. (2024) reported 16–20% DOI accuracy; we observe 75.5%, a substantial improvement attributable to search tools. Aljamaan et al. (2024) found that standalone LLMs scored at the maximum hallucination level on their weighted Reference Hallucination Score, while retrieval-augmented tools scored near zero—a pattern corroborated by the large gap between our search-enabled baselines and the non-search results reported in earlier work. Unlike these prior studies, our evaluation is version-aware: ground truth accommodates preprint-to-published metadata

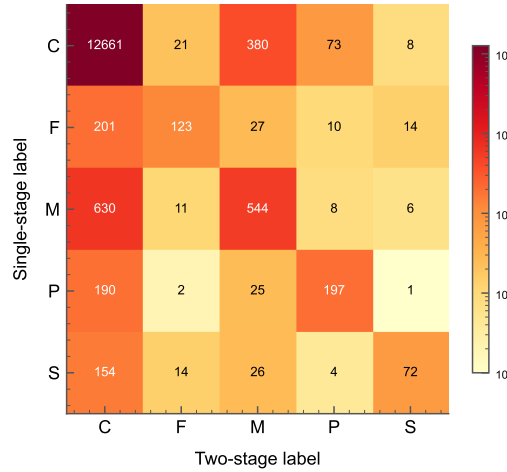


Figure 18: Label transition matrix between single-stage (rows) and two-stage (columns) integration, aggregated over GPT-5 and Claude. The dominant off-diagonal transitions involve missing (M) labels: 630 fields are corrected from M to C by two-stage; 380 regress from C to M. Log-scale coloring accommodates the $C \rightarrow C$ cell (12,661).

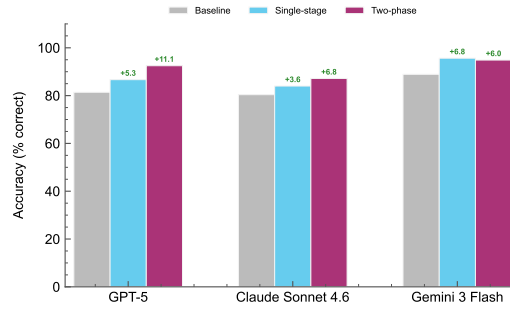


Figure 19: Field-level accuracy by model and integration strategy (baseline, single-stage clibib, two-stage clibib). Single-stage results are shown for GPT-5 and Claude only; Gemini lacks single-stage support. Two-stage integration improves all three models over baseline.

changes, reducing false negatives for fields such as venue, volume, and pages that shift across versions.

H.2 Citation count and memorization

Niimi (2025) found a log-linear correlation ($r = 0.75$) between citation count and metadata accuracy, with a memorization threshold near 1,000 citations. Our citation-tier stratification provides converging evidence under search-enabled conditions: popular papers (median citations $> 1,000$) achieve the highest fully correct rates, while recent and low-citation papers degrade sharply—consistent with Niimi’s redundancy-driven account, though web search partially compensates for lower parametric familiarity.

H.3 Comparison with unmediated database export

The comparison with Szeider (2026) is instructive. Szeider’s unmediated DBLP export achieves 82.7% perfect entries (all fields correct); our baseline search-enabled models achieve 50.9%. This gap reflects the fundamental difference between deterministic database export and LLM-mediated generation. With two-stage clibib augmentation, our models reach

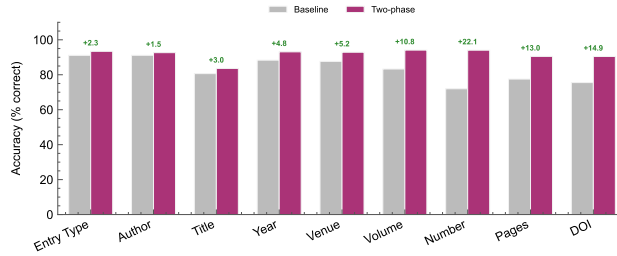


Figure 20: Per-field accuracy before (baseline) and after two-stage clibib revision, averaged across all three models. Number (+22.1 pp), doi (+14.9 pp), and pages (+13.0 pp) show the largest gains.

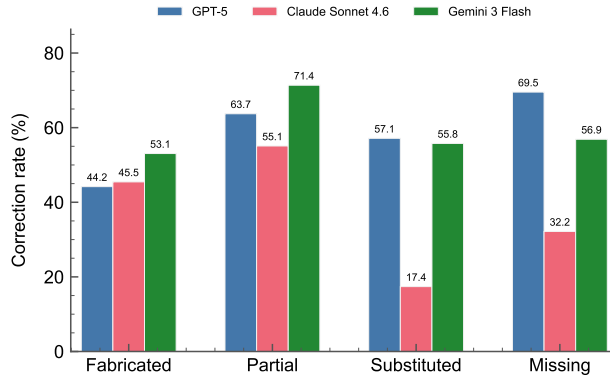


Figure 21: Two-stage correction rate by original error type, per model. GPT-5 shows the highest correction rates for partial and missing errors; Gemini corrects partial errors most effectively (71.4%).

78.3% fully correct entries, narrowing this gap substantially. The architecture of the LLM–database interaction also matters: our ablation shows that separating retrieval from revision (two-stage) reduces regression to 0.8%, compared with 4.8% for single-pass tool loops where the model interleaves search and generation. This reinforces Szeider’s finding that how the model accesses authoritative data—not just which data source is used—determines output quality.

More broadly, fully deterministic pipelines that resolve a known DOI through the Cross-Ref API to publisher-deposited metadata can achieve near-perfect field accuracy for DOI-bearing papers. However, these pipelines require a structured identifier as input, which is unavailable in the knowledge-retrieval scenario we study: users provide natural-language descriptions, not DOIs or arXiv IDs. clibib occupies a middle ground, accepting natural-language queries and resolving them to authoritative records via the Zotero Translation Server (which queries CrossRef and DBLP); a direct comparison isolating each upstream data source is future work.

H.4 Structural hallucination

Boudourides (2026) frames bibliographic errors as instances of *structural hallucination*: systematic distortion of relational architecture that remains invisible to sentence-level fluency checks. Without retrieval tools, gpt-4.1-mini achieved DOI F1 = 0.01 and omitted 91.9% of ground-truth citations; our search-enabled models achieve 75.5% DOI accuracy, confirming that retrieval substantially closes this gap. Field-level metadata errors—particularly in DOIs and venue identifiers—propagate into citation networks, corrupting link structure and distorting downstream bibliometric analyses. Our results support Boudourides’ call

for pre-ingestion verification when LLM-generated bibliographic records enter scholarly infrastructure.