HALLMARK: HOW TO DIAGNOSE LLM CITATION VERIFIERS AND WHEN TO DEPLOY THEM

Anonymous authors
Paper under double-blind review

ABSTRACT

GPTZero found 53 NeurIPS 2025 accepted papers with hallucinated citations (Shmatko et al., 2026; Ansari, 2026), and venues now screen submissions with automated reference checkers. Rule- and Large Language Model (LLM)-based verifiers are emerging, but many are released without a shared evaluation and each flags a different kind of error, so it is unclear which to trust or when to deploy them. We address this with HALLMARK (**Hall**ucination bench**mark**): 2,526 BibTeX entries spanning 14 hallucination types, with per-type and diagnostic sub-test labels and further temporal, cross-domain, and real-world evaluation splits. We evaluate zero-shot and tool-augmented LLMs, a DOI-lookup baseline, and our own co-designed rule-based verifier, `bibtex-updater`. Verifiers can miss hallucinations and wrongly flag correct papers, so every flag still needs human review. The false positive rate (FPR) determines the review effort and thus limits deployment. The tested models exhibit an order-of-magnitude spread in FPR; agentic lookups buy recall but inflate FPR, and neither smaller models nor prompt optimization reach the frontier's operating point. A two-stage cascade with `bibtex-updater` avoids that inflation, providing an upper bound, given the co-design. These rankings, however, depend on what counts as a hallucination: re-scoring the types that describe a real work cited inaccurately changes which verifier ranks first, and venues have not agreed on a definition. We call for a community definition and release per-type labels and a re-scoring protocol so that any agreed definition can be scored against every verifier.

1 INTRODUCTION

Hallucinated citations—fabricated references generated by language models—have become a prevalent threat to scientific publishing, especially because they often *seem* plausible. Shortly after NeurIPS 2025, GPTZero audited the 4,841 accepted papers and found 53 with fabricated citations that had passed peer review (Shmatko et al., 2026; Ansari, 2026), and independent audits report hundreds more across venues and a sharp rise since 2021 (Xu et al., 2026; Sakai et al., 2026a; Bienz et al., 2026). Venues now screen submissions for them: ICLR 2026 ran an automated reference checker, desk-rejected papers with confirmed hallucinations, and reported a significant false-positive rate (ICLR, 2026).

Verifiers have multiplied, from DOI resolvers to multi-database cross-referencing systems and LLM-based checkers. Many are released without a shared evaluation, and often focus on different error types, ranging from records that cannot be found to mismatched authors, venue or year. It is unclear which to trust, or how to deploy them.

To enable comparisons and to provide practical guidance on deployment for authors and venues, we release HALLMARK, comprising 2,526 BibTeX entries spanning 14 hallucination types across three difficulty tiers, with taxonomy labels and six diagnostic sub-tests. We evaluate a DOI-lookup baseline, frontier and small LLMs zero-shot, tool-augmented agents, and our own rule-based verifier `bibtex-updater`, which was co-designed with the taxonomy and is therefore reported as an upper bound rather than ranked. We score every verifier under a single protocol (Fig. 1) and rely on both a held-out test set and four additional evaluation-only splits that add recent papers, a second domain, and authentic ChatGPT citations.

Scope. HALLMARK targets citation *metadata*—detecting fabricated, inconsistent, or nonexistent bibliographic records—not claim-level hallucination, which needs full-text analysis.

Current citation verifiers are not perfect; thus, human review is required, and *the false-positive rate, not recall, becomes the bottleneck for deployment*, as we show in our positive predictive value (PPV)

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

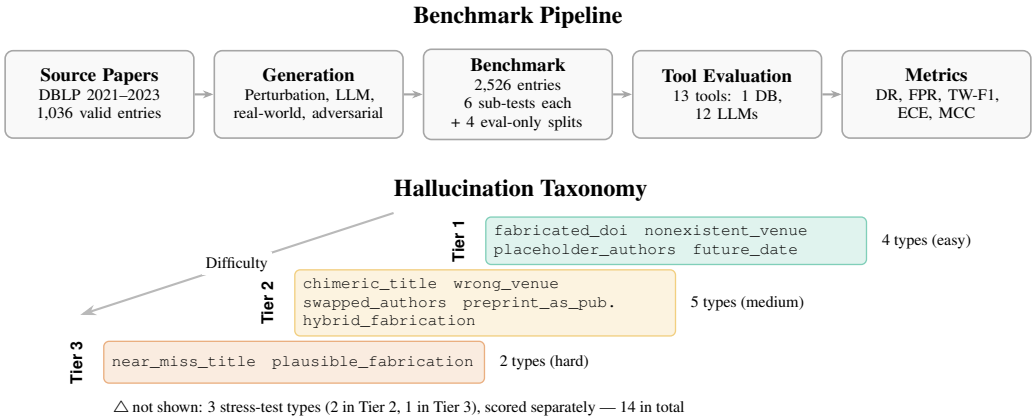

Figure 1: **Overview of HALLMARK. Top:** The benchmark pipeline: real papers sourced from DBLP are transformed via perturbation, LLM generation, and real-world collection into 2,526 annotated BibTeX entries, each with six diagnostic sub-tests; four evaluation-only splits (1,349 entries) are scored separately (Tab. 5). Thirteen full-coverage verification tools (1 citation-database, 12 zero-shot LLMs) are evaluated using tier-weighted metrics. **Bottom:** The three-tier hallucination taxonomy with the 11 *main* types, ordered by the verification effort required to detect them; 3 further stress-test types are evaluated separately and listed in Tab. 2, for 14 in total.

analysis in § 4.4, with FPR spanning an order of magnitude (0.043 to 0.702). We observe and analyze three failure modes (Tab. 20): (i) agentic lookups buy recall but inflate FPR; (ii) precision collapses at realistic prevalence; and (iii) most LLMs by default over-flag papers published after their training cutoff. Some of these failure modes can be mitigated: a two-stage cascade with `bibtex-updater` avoids inflating FPR by passing only uncertain entries to an LLM. It catches nearly all hallucinations at a low false-positive rate, which we treat as an upper bound given the co-design. A prompt that reminds the model of its training cutoff (i.e., cutoff-aware) reduces over-flagging of recent papers, but mostly through abstention; mitigating it selectively is left for future work. Prompt optimization (Agrawal et al., 2026) also has limits, especially for smaller models, which maintain high FPR.

Verifier performance and ranking depend on the—perhaps surprisingly ambiguous—definition of a hallucination. For example, a real work cited with the wrong venue is not fully correct, but is that a hallucination? Existing verifiers address this differently in what they check, and venues have not (yet) reached consensus. We therefore urge the community to have more discussion on definitions of hallucinated citations; meanwhile, we release per-type labels and a re-scoring protocol so that any agreed definition can be scored against every verifier.

Our **contributions** are:

1. **A benchmark and taxonomy** for citation-hallucination detection, with per-type and sub-test labels and temporal, cross-domain, and real-world splits (§ 3).
2. **A failure-mode analysis** of citation verifiers under one protocol, showing that the false-positive rate decides deployability (§ 4.4).
3. **A re-scoring protocol** that lets any definition of a hallucinated citation be applied to every verifier.

2 RELATED WORK

Citation hallucination in the LLM era. LLM hallucination (Ji et al., 2023; Huang et al., 2025) manifests distinctly in citations (Alkaissi & McFarlane, 2023; Agrawal et al., 2024; Walters & Wilder, 2023); venue-level audits document scale—GhostCite (Xu et al., 2026) on 56K papers, HalluCitation (Sakai et al., 2026a) on ACL/NAACL/EMNLP, the HPC-venue study of Bienz et al. (2026), and CITADEL's audit of 2.5 million PubMed Central papers (Topaz et al., 2026)—measuring *prevalence*, not detection-tool performance. Detection tools—HaRC (HaRC Contributors, 2024), verify-citations (verify-citations Contributors, 2025), RefChecker (Hu et al., 2024), CheckIfExist (Abbonato, 2026)—report incompatible metrics on disjoint test data, preventing direct comparison; CheckIfExist (a CrossRef/Semantic-Scholar/OpenAlex cascade verifier) ships without dataset or metrics, and RefChecker requires full-text manuscript context. Concurrent work (Rao & Callison-Burch, 2026; Rao et al., 2026; Sakai et al., 2026b) evaluates LLM-as-citation-generator accuracy or releases lightweight checkers. The closest work to HALLMARK is CiteAudit (Shi et al., 2026),

which likewise benchmarks LLMs and commercial verifiers on a human-validated/annotated dataset via a multi-agent pipeline. HALLMARK differs by focusing on understanding failure modes by rich diagnostics. It provides a 14-type, tier-stratified error taxonomy and post-cutoff evaluation splits that expose training-cutoff effects, so it reports *which* metadata failures each tool catches. Adjacent concurrent detection work (Rao et al., 2026) targets citation-URL liveness on DRBench (the stale-vs-hallucinated distinction), a scope orthogonal to HALLMARK's BibTeX-metadata cross-consistency checks. § C.7 details boundary cases of audit-vs-benchmark categorization.

Hallucination detection benchmarks. General-purpose hallucination benchmarks (Li et al., 2023; Ravichander et al., 2025) target factual claims, not metadata; adjacent work—FactScore (Min et al., 2023), RARR (Gao et al., 2023), SciFact (Wadden et al., 2020)—does not evaluate BibTeX integrity (Digital Object Identifier (DOI)/title/author/venue cross-consistency against external databases). Concurrent generator-side benchmarks such as HalluHard (Fan et al., 2026) measure whether LLMs produce *claims* grounded by retrievable inline citations in multi-turn dialogs. None evaluates the verifiers that check citation metadata. HALLMARK's design borrows from established benchmarks: temporal segmentation, as in SWE-bench (Jimenez et al., 2024) and LiveCodeBench (Jain et al., 2025), separates recall from verification, and multi-criteria sub-tests, as in HumanEval (Chen et al., 2021), attribute each failure to a field (§ C.7).

3 THE HALLMARK BENCHMARK

3.1 HALLUCINATION TAXONOMY

We define 14 citation hallucination types—11 empirically grounded and 3 stress-test types—organized into three difficulty tiers based on the verification effort required (Tab. 2, in the appendix). **Tier 1 (Easy)** hallucinations are settled by one check, a single lookup (`fabricated_doi`, `nonexistent_venue`) or none at all (`placeholder_authors`, `future_date`). **Tier 2 (Medium)** hallucinations require cross-referencing multiple metadata fields: e.g., `chimeric_title` pairs real authors with a fabricated title, `wrong_venue` assigns a paper to the wrong conference, and `hybrid_fabrication` uses a real DOI whose resolved record does not match the BibTeX metadata. **Tier 3 (Hard)** hallucinations require deep verification or semantic reasoning: `near_miss_title` entries differ by one or two words from a real paper; `plausible_fabrication` entries are entirely invented but realistic; and `arxiv_version_mismatch` entries cite a preprint with wrong venue and shifted year. The tiers rank the effort a human verifier needs, not difficulty for an LLM.

The taxonomy is grounded in real incidents: we derived it from hallucinated citations in the NeurIPS 2025 incident and related audits (Xu et al., 2026; Sakai et al., 2026a), then stress-tested it with adversarial brainstorming of failure modes existing tools might miss. Of the 108 real-world hallucinated citations in the benchmark, the 72 we analyzed in detail map cleanly to a single type 79% of the time, with the rest assigned to their highest applicable tier (§ A.2). Three types—`merged_citation`, `partial_author_list`, `arxiv_version_mismatch`—we could not trace to a documented incident, so we treat them as stress-test types. They appear in every split (18.3% of hallucinated entries on `dev_public`, 20.2% on `test_public`); per-type labels let each type be scored separately. Perturbation rules, generation prompts and the collection procedure are in § A.3 and § B.2.3.

3.2 DATASET CONSTRUCTION

The dataset contains two classes of entries: *valid* references scraped from DBLP and *hallucinated* references built with the four methods below.

Valid entries. We scraped BibTeX records from DBLP (dblp Team, 2025) for papers published at major ML venues (NeurIPS, ICML, ICLR, AAAI, CVPR) between 2021 and 2023, plus 35 arXiv preprints from February 2026 (22 `dev_public`, 13 `test_public`) as a recency probe. Each entry was verified by confirming DOI resolution where a DOI is present (46% of valid entries), title existence in at least two databases, and author-venue consistency. This retained 1,036 valid entries.

Hallucinated entries. We generated hallucinated entries using four methods: *Systematic perturbation*: modifying specific fields of valid entries to produce targeted hallucination types (e.g., replacing a DOI with a non-resolving one for `fabricated_doi`, swapping author lists between papers for `swapped_authors`). *LLM generation*: prompting language models to generate plausible but fictional references for types requiring coherent fabrication (`plausible_fabrication`,

chimeric_title). *Adversarial crafting*: manually constructing entries designed to evade specific detection strategies. *Real-world collection*: harvesting actual hallucinated citations from published papers identified in audits. An evaluation-only supplement of 341 authentic ChatGPT-generated citations (172 valid / 169 hallucinated), hand-verified by Walters & Wilder (2023) across 42 multidisciplinary topics, adds real-world hallucinations (§ C.6).

Labeling. Labels are assigned deterministically by the generation pipeline; LLM-generated entries are also filtered against bibliographic databases, and real-world and adversarial entries receive manual review (§ A.3).

Splits and release. Every hallucinated entry is built by modifying some real *base paper*. Across the two public splits those base papers do not overlap: no hallucinated entry on one split is built from a paper that appears as a valid entry on the other. test_hidden and stress_test do not have this property: 109 of their 244 and 68 of their 121 hallucinated entries modify a paper that is a valid entry in public data. Most types on dev_public and test_public carry ≥ 30 instances, enough for meaningful per-type comparison; test_hidden spreads 244 hallucinated entries over 14 types and does not clear that floor, so its per-type intervals are wider (statistical power in § B.1). Every entry passes automated validation, the url field is blinded to avoid using URL resolution as a shortcut for verification, unique marker tokens in the metadata expose memorization if a model reproduces one, and types are roughly balanced to maximize per-type power (§ A.3). The core splits are dev_public (1,119 entries), test_public (831), test_hidden (454), and stress_test (122); per-split counts and the four evaluation-only splits are in Tab. 5. Entries carry a temporal segment tag (*pre-2023*, *2023–2024*, *2025+*), used for the post-cutoff analysis in § 4.4 and § F.

Sub-test design. To understand what is hallucinated and *why*, each entry decomposes citation validity into six independently verifiable dimensions, each scored True/False/N/A: **DOI resolves**, **Title exists** in bibliographic databases, **Authors match** the identified paper, **Venue real** and correctly attributed, **Fields complete**, **Cross-DB agreement**.
The six sub-tests also make an entry's defect *count* observable: one type can span several failed sub-tests, and multi-defect entries are the norm in the real-world subset, whereas under perturbation they are rare. Detection by defect count is in § C.5.3.

4 Experiments

4.1 Metrics

Our comparison focuses on two *prevalence-independent* metrics, ensuring they remain comparable across splits with differing class ratios. **Detection rate (DR)** is recall on the hallucinated class; **FPR** is the fraction of valid entries wrongly flagged. **Tier-weighted F1 (TW-F1)** weights each hallucinated entry by its tier (1/2/3), rewarding the detection of harder hallucinations. **Matthews Correlation Coefficient (MCC)** reads all four confusion-matrix cells; it moves with prevalence by ≤ 0.03 between our splits (54.2% vs. 62.5% hallucinated). **Expected Calibration Error (ECE)** (Naeini et al., 2015) measures how well a tool's confidence tracks its accuracy; an ECE below 0.05 is excellent and above 0.2 unreliable. Formal definitions are in § B.1; per-tier and per-type breakdowns in §§ C.4 and 4.3.

4.2 Evaluated tools

We evaluate thirteen independent full-coverage tools (i.e., returning a verdict on every entry, neither co-designed with the benchmark nor given tool access) on dev_public: a DOI-only lower bound and twelve zero-shot LLMs from OpenAI, Anthropic, Google, Meta, Mistral, DeepSeek and Qwen (Tab. 1; model IDs in § B.2.4), including a later-cutoff GPT-5.4 control (Aug 2025 cutoff) for § 4.4. All LLMs use the same zero-shot prompt at temperature 0 and seed 42 with one API call per entry; DeepSeek-R1's chain-of-thought raises its median latency to $\sim$19 s against 1–10 s for the single-pass cohort (Tab. 19). On top of these, four *agentic* variants (§ B.2.4) allow up to five database lookups per entry; prompted to cross-reference databases, the model tends to flag an entry as soon as one lookup returns no match—the disposition behind the inflated false-positive rate we report in § 4.4. HaRC (HaRC Contributors, 2024) and verify-citations (verify-citations Contributors, 2025) are excluded from Tab. 1: neither returns its own verdicts at benchmark scale, so any row for them would report the shared pre-screening layer (§ B.2.2). The DOI-only baseline only verifies each entry's DOI against the DOI.org resolver. The optional pre-screening layer (DOI format, year bounds, author heuristics) is applied to the database-lookup tools and the cascade, not to the LLM verifiers, so those

Table 1: **Main results on `dev_public`. Bold** marks the best *point estimate* among independent full-coverage tools (DOI-only and the zero-shot LLMs, i.e., neither co-designed nor tool-augmented; see the *Scoring conventions* paragraph, § 4.3). The shaded *co-designed* block was developed alongside the benchmark taxonomy and is a reference *upper bound excluded from ranking*; no co-designed cell is bolded (§ G.2). The zero-shot block is sorted by FPR ascending. **Coverage** (Cov.) is the fraction of entries a tool decides about, scored under the *Scoring conventions*. ΔFPR is the cross-split shift `test_public` − `dev_public`. FPR cells are shaded: blue marks the low-FPR band (FPR $\leq$.13), red marks over-flagging operating points (FPR $\geq$.41); the gray co-designed block is excluded from shading as from ranking. The three cascade rows count entries neither stage resolves as hallucinated (Tab. 22); they share one Stage 1 (Stage 1: `bibtex-updater`; Stage 2: an LLM; § 4.5), and flagging rather than dropping the entries Stage 2 leaves undecided changes no cascade's FPR. `bibtex-updater` is scored over the full split with abstentions resolved to VALID; under the LLM scoring rule (abstentions excluded; § E.2) it reaches DR 0.984 / FPR 0.056 at coverage 0.87. *Always-call*: `bibtex-updater` runs on every entry and its output is given to the LLM in the prompt.

	Performance					Calibr.	Coverage	Robustness
Tool	**DR ↑**	**FPR ↓**	**F1 ↑**	**MCC ↑**	**TW-F1 ↑**	**ECE ↓**	**Cov. ↑**	**ΔFPR ↓**
Citation-database tools								
DOI-only	0.218	**0.043**	0.347	0.253	0.268	0.082	1.00	− 0.001
Zero-shot LLMs (sorted by FPR)								
Gemini 2.5 Pro	0.476	0.050	0.627	0.473	0.609	0.297	0.97	+0.009
Gemini 2.5 Flash	0.500	0.100	0.631	0.429	0.628	0.265	0.99	+0.006
Llama 4 Maverick	0.614	0.146	0.707	0.476	0.709	0.176	1.00	+0.020
Claude Opus 4.7[§]	0.906	0.154	0.890	0.755	0.913	0.052	1.00	+0.051
Claude Sonnet 4.6[§]	**0.916**	0.161	**0.893**	**0.760**	**0.917**	**0.041**	0.98	+0.026
GPT-5.4 (zero-shot)	0.767	0.228	0.783	0.538	0.807	0.202	1.00	− 0.004
Mistral Large	0.716	0.250	0.742	0.465	0.765	0.229	0.99	+0.032
GPT-5.1 (zero-shot)	0.837	0.411	0.766	0.442	0.822	0.190	1.00	+0.069
Qwen3-235B	0.860	0.533	0.744	0.358	0.821	0.279	1.00	+0.082
Qwen3-VL-235B	0.860	0.551	0.740	0.342	0.818	0.286	1.00	+0.077
DeepSeek-R1	0.896	0.623	0.739	0.324	0.825	0.238	0.98	− 0.303[¶]
DeepSeek-V3.2	0.911	0.702	0.727	0.268	0.821	0.316	1.00	+0.026
Agentic (tool-use; up to 5 tool calls per entry)								
GPT-5.1 + CrossRef/OpenAlex/arXiv	0.967	0.478	0.816	0.558	0.892	0.175	1.00	+0.080
GPT-5.1 + bibtex-updater (agentic; tool optional)	0.980	0.470	0.824	0.584	0.900	0.125	1.00	− 0.114
Sonnet 4.6 + bibtex-updater (agentic; tool optional)[‡]	0.990	0.431	0.841	0.630	0.913	0.118	1.00	− 0.088
Co-designed (reference upper bound; see § G.2)								
bibtex-updater (v1.2.0)	0.820	0.051	0.880	0.768	0.892	0.395	0.86	+0.029
Cascade: bibtex-updater → Sonnet 4.6 (§ 4.5)[♯]	0.990	0.037	0.980	0.955	0.986	0.117	1.00	+0.021
Cascade: bibtex-updater → GPT-5.1 (§ 4.5)	0.974	0.039	0.970	0.935	0.976	0.091	1.00	+0.019
Cascade: bibtex-updater → GPT-5.4 (§ 4.5)[♯]	0.990	0.039	0.979	0.953	0.985	0.097	1.00	+0.019
GPT-5.1 + bibtex-updater (always-call; output in prompt)	0.843	0.144	0.856	0.698	0.872	0.078	0.93	+0.112

[¶]DeepSeek-R1's ΔFPR of −0.303 is measured over the 651 `test_public` entries it answered, the other 180 being API error fallbacks (Tab. 9). [§]Collected 2026-09-02. [‡]Contains Sonnet 4.6 from the earlier 2026-05-05 collection. [♯]Collected 2026-09-21 on the Stage 1 shared by all three cascades. Comparing a [§] row against a [‡] row crosses endpoint snapshots, and the drift (Tab. 29) can change the reading. Unmarked rows: 2026-05-05.

rows include their verdicts (§§ D.1 and 4.4). We label `bibtex-updater` *co-designed* because its development overlapped the taxonomy: the typed sub-tests partly mirror the tool's verification stages, so it may score better here than on a novel hallucination distribution. We read its row as an unranked *upper-bound reference* (§ G.2).

4.3 Main results

Scoring conventions. Rows in Tab. 1—which reports `dev_public`—are ranked by point estimate, with a paired test only where per-entry predictions allow (§ B.1); an LLM's UNCERTAIN verdicts are excluded from the DR/FPR/F1 triple, while `bibtex-updater` resolves its abstentions to VALID before scoring (§§ B.1 and E.2); predictions were collected 2026-05-05 apart from the two rows marked [§] in Tab. 1, from a later collection (2026-09-02), and hosted-LLM rows are dated snapshots subject to endpoint drift (§ 5).

LLMs span a wide recall–precision spectrum. The FPR ordering of Tab. 1 (DR vs. FPR tradeoff in Fig. 2) runs from models that flag rarely (Gemini 2.5 Pro) to models that flag most entries (DeepSeek-V3.2). The two later-cutoff Anthropic models give the best balance—Opus 4.7 and Sonnet 4.6 have the highest F1 and the best calibration of the independent tools, at a middling FPR. Their F1

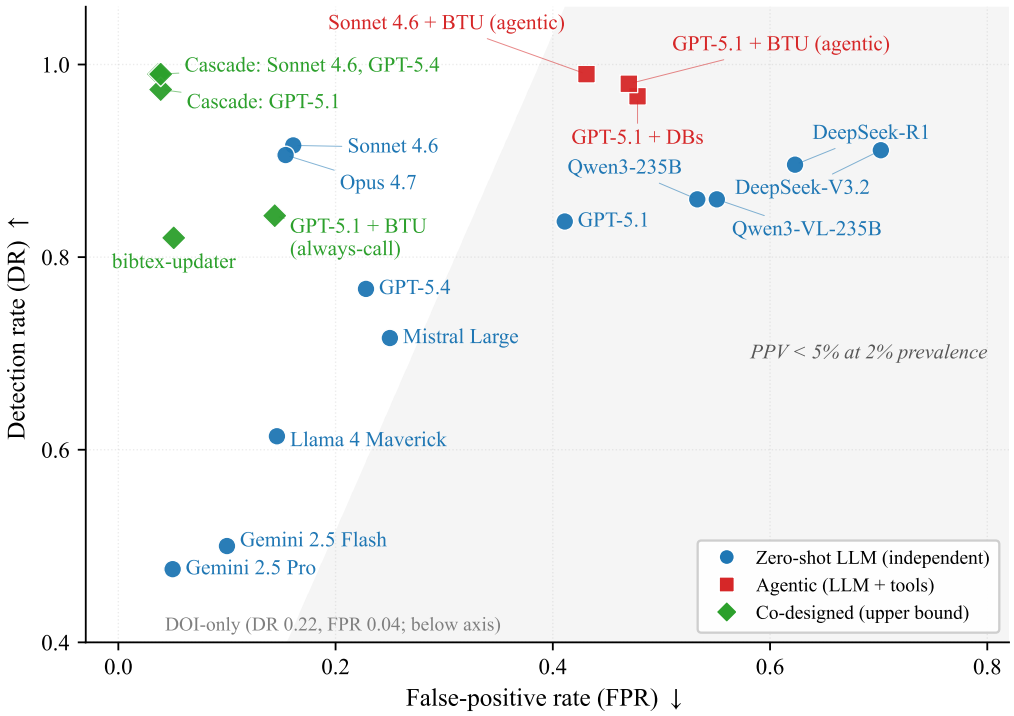

Figure 2: **DR–FPR operating points on `dev_public`.** Each point is a tool's operating point, the (FPR, DR) pair it reaches, one per row of Tab. 1. BTU: `bibtex-updater`; DBs: CrossRef/OpenAlex/arXiv. *Agentic* (red): the LLM may call BTU or the DBs as a tool, up to five calls per entry; *always-call* (green): BTU runs on every entry and its output is given to the LLM. In the shaded region, more than 19 out of every 20 flags are false alarms at a venue-realistic rate of 2% hallucinated citations (§ 4.4), so high recall alone does not make a tool usable. The co-designed cascades reach high recall at low FPR (an upper bound).

gap is only 0.3 pp (paired bootstrap p=0.74, § B.1), and it reverses direction between the May and September runs: Opus 4.7 leads on the 2026-05-05 run, Sonnet 4.6 on the 2026-09-02 one (Tab. 29). Neither model beats the other on both axes—Opus 4.7 has the lower FPR (.154 vs. .161), Sonnet 4.6 the higher recall (.916 vs. .906)—and on the earlier 2026-05-05 collection both sat at the low-FPR end (.072, .127; Tab. 29). Sonnet's extra recall is also partly over-flagging: of the 27 `dev_public` papers the ground-truth audit reclassified as real (§ 3.2), Sonnet flagged 19 as hallucinated and Opus 8. Both clear GPT-5.1 and the open-weight models that flag most entries, and GPT-5.4 (cutoff Aug 2025) sits between them at lower FPR and comparable recall, consistent with § 4.4.
§ E.2 examines whether `bibtex-updater`'s low FPR depends on abstention. Because DR and FPR are prevalence-independent, we compare on them, and on F1 only within a split.

Per-type analysis. Per-type detection rates (heatmap in Fig. 3; full table in Tab. 13) show the same pattern across models: most LLMs catch nearly every Tier-1 hallucination, while `swapped_authors` and `near_miss_title` are the hardest types for almost every model. GPT-5.1 reaches at least 74% on 9 of 11 main types and is perfect on `nonexistent_venue` and `placeholder_authors`. Its two weak types, `swapped_authors` and `near_miss_title`, both turn on exact bibliographic knowledge and map to the *Authors match* and *Title exists* sub-tests respectively (§ 3.2; Tab. 2).

Small models. As not every researcher can access the largest (and, often, costliest) models, we investigate whether a small local model can be a good substitute. We test on Claude Haiku 4.5 and four Qwen3 open-weight models at 4, 8, 14 and 32B, under the same protocol as Tab. 1, and report the details in § C.3. Haiku performs substantially worse than the tested Anthropic Sonnet and Opus models, and the smaller open-weight models flag more than 95% of the split as hallucinations, yielding a high DR carrying no information: at venue-realistic prevalence their PPV is indistinguishable from

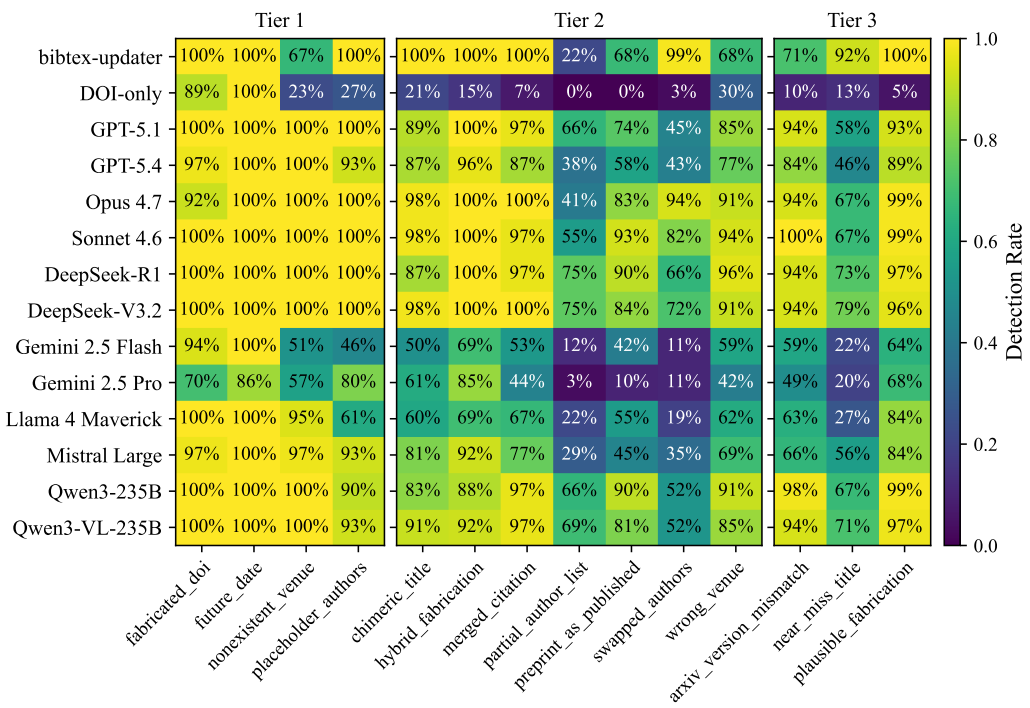

Figure 3: **Per-type detection rates across all full-coverage tools.** The 14 types are grouped into three panels by difficulty tier (Tab. 2). HaRC and verify-citations are excluded because neither produces its own verdicts at benchmark scale (§ B.2.2). `swapped_authors` and `near_miss_title` remain the hardest types across the cohort, and the tier grouping makes the pattern visible: most LLMs catch nearly every Tier-1 hallucination, while Tiers 2–3 separate the cohort; tiers rank the effort a human verifier needs, not difficulty for an LLM.

the base rate (§ E.1), so nearly every entry requires manual checks. Thus, cheap inference does not make deployment cheap.

Cross-split robustness. The held-out `test_public` (831 entries) tells a complementary story to the `dev_public` results (ΔFPR in Tab. 1; full results in § C.1). Among standalone tools with DR $\geq$0.80, `bibtex-updater` keeps the lowest false-positive rate on both splits (0.051 on `dev_public`, 0.080 on `test_public`), which we read as an upper bound given the co-design; out of domain its FPR stays low but its coverage drops (§ G.2).

Prompt sensitivity. A prompt-sensitivity ablation on 150 `dev_public` entries (four prompt variants, four models; § H.2) moves the same model's absolute FPR by 10–37 pp on wording alone, while the model ranking barely moves (mean pairwise Spearman rank correlation ρ=0.90): absolute FPR values depend on the prompt, the model ordering does not.

Takeaway. Even the highest-recall independent model (Sonnet 4.6) misses $\sim$8% of hallucinations, and its misses concentrate on subtle types: `near_miss_title` and `swapped_authors` are its weakest, both demanding exact bibliographic knowledge (Tab. 13). Among independent tools, Sonnet 4.6 and Opus 4.7 give the best F1 and calibration at a middling FPR, a ranking that survives a prompt-paraphrase ablation (§ H.2).

4.4 Failure modes

Three failure modes recur across the cohort (Tab. 20); we diagnose each in turn.

Agentic lookups give diminishing returns (failure mode (i)). Five tool calls push GPT-5.1's recall past `bibtex-updater`'s, but its false-positive rate rises to roughly nine times `bibtex-updater`'s, because the prompted model—cross-referencing against CrossRef/OpenAlex/arXiv—tends to flag an entry as soon as one of them returns no match. The inflation is a property of the harness—its naive, single-stage use of the lookups—not of the base model or of tool access itself: swapping Sonnet 4.6 for GPT-5.1 reproduces the profile within $\leq$5 pp

on every metric, and Sonnet's FPR rises $\sim 3.4\times$ over its zero-shot baseline from the same collection, while GPT-5.1's already-high zero-shot FPR barely changes. This any-no-match policy fails because CrossRef, OpenAlex, and arXiv have partial, non-overlapping coverage: a real paper is routinely missing from one of them, so single-source absence gets read as fabrication. A better alternative is to flag only on *consensus* absence across sources (or on positive metadata inconsistency) and to route single-source absence to UNCERTAIN/human review (§ E.2); § 4.5 measures that policy. A deterministic re-aggregation over three of the harness's four databases confirms improvement (§ D), so failure mode (i) characterizes a retrieval-augmented pattern, not an inherent shortcoming of agentic systems. Tool access therefore trades precision for recall: agentic harnesses flag an entry as soon as one database misses it.

Precision at low prevalence (failure mode (ii)). Precision (PPV, the fraction of flags that are genuine hallucinations) falls at low prevalence by Bayes' rule (§ E.1). Our analysis focuses on how verifiers compare, given the wide range of FPR in Tab. 1, from 0.043 to 0.702. At a venue-realistic $\sim 2\%$ base rate that becomes a $\sim 6\times$ gap in precision: the best verifiers catch one true hallucination per 6–10 flags, the most aggressive fewer than one in 35, even at $\geq 87\%$ detection (per-tool PPV in Tab. 23, sweep in Tab. 24). Calibration does not simply follow FPR: Sonnet 4.6 and Opus 4.7 are the best-calibrated zero-shot LLMs (ECE 0.041, 0.052) at a middling FPR, while the lowest-FPR model, Gemini 2.5 Pro, is among the worst calibrated (0.297). The co-designed `bibtex-updater` has a low FPR because it flags conservatively rather than abstaining, and reaches one-in-four precision (§ E.2). The optional pre-screening layer adds ~ 5 pp overall (~ 18 pp on Tier 1); we report numbers with and without pre-screening. Every metric from which abstentions are excluded is paired with its coverage and an aggressive re-scoring that scores every abstention as HALLUCINATED (§ E.2), so deferring hard cases cannot improve low precision.

Prevalence (the base rate of hallucinated entries) moves deployability while leaving the ranking fixed. Sweeping the base rate over 1–5% leaves the PPV ordering essentially fixed (Tab. 24), because prevalence cancels from pairwise comparisons. Thus, the DR/FPR ratio determines the ranking; however, the base rate affects *absolute* precision: at a 2% base rate even the best independent verifier (Gemini 2.5 Pro) reaches only $\sim 16\%$ PPV. Still, depending on prevalence, the cost ratio $c_{\text{FN}}/c_{\text{FP}}$, and reviewer capacity, it might be worth ranking based on recall. We discuss the different deployment considerations and regimes in § E.3.

> **Takeaway.** Whether a verifier is deployable depends mostly on its FPR, as checking (falsely) flagged entries requires human oversight. At venue-realistic prevalence, precision is set by the ratio of detection rate to false-positive rate. In our experiments, FPR spans an order of magnitude, which substantially influences precision (PPV), and thus, whether deployment is feasible.

Temporal robustness (failure mode (iii)). A separate 448-entry supplement disjoint from the 2021–2023 corpus (300 valid 2024–2025 DBLP entries, 148 hallucinated; § F.1) shows LLM verifiers degrade sharply beyond their training cutoff. The earlier-cutoff models default to "flag everything unfamiliar": F1 falls to ≈ 0.55 as false positives inflate, with GPT-5.1's FPR alone rising $41.1\% \rightarrow 75.9\%$. DeepSeek-R1's detection rate is not evaluable: API error fallbacks took 151 of its 448 entries, all 148 hallucinated ones among them; on the valid entries it answered, FPR$=0.856$ (full panel: Tab. 28). As with every absolute FPR here, these values depend on the prompt, so we read them through the model ordering (*Prompt sensitivity*, § 4.3). Retrieval need not help: databases also lag on recent papers (§ F.2).

Later-cutoff resistance. Only the two latest-cutoff Anthropic models hold their false-positive rate near in-distribution levels (Sonnet 4.6 12.0%, Opus 4.7 7.3%); Gemini 2.5 Pro and GPT-5.4 (Tab. 31) over-flag only moderately, and the remaining eight degrade sharply to 59.5–88.7% (the "over-flagging cluster"; Tab. 28). The two Anthropic models' dev$\rightarrow$temporal FPR barely moves (-0.7 and $+0.1$ pp). Both figures in that comparison come from the May experiment, since a shift across the cutoff is only interpretable within one endpoint snapshot; the September re-run of these two models reported in Tab. 1 puts their `dev_public` FPR higher (.161, .154), a drift larger than the temporal shift itself (we provide more details about endpoint drift in Tab. 29). We report this as a *descriptive* pattern, not a calibration result, because the design cannot cleanly separate two explanations: the valid pool is scraped from DBLP and these are the latest-cutoff models in the cohort, so a low FPR on a valid entry is indistinguishable from training-data *recall* of that same DBLP record; the risk of such contamination—benchmark entries appearing in a model's training corpus—grows with capability, as stronger models train on larger, newer corpora (Rank et al., 2026).

Two controls, a recall probe and a third-provider late-cutoff model, DeepSeek-V4-Pro, add nuance but do not eliminate the memorization reading; with two models from one provider and a single third-party control, it remains a preliminary signal (§ F.4).

Takeaway. The two latest-cutoff models keep their post-cutoff FPR low while earlier-cutoff models degrade sharply, and the one third-provider late-cutoff control we ran (DeepSeek-V4-Pro) stays close to Sonnet 4.6. For Opus 4.7, a recall probe points to calibration rather than memorization. We read this as a preliminary descriptive signal, not a causal finding: it rests on a single third-party control, the Anthropic runs have not yet been reproduced on Anthropic's native API (only through the OpenRouter mirror), and the recall probe relies on whether a model can name the authors from the title alone.

4.5 Stage-2 diagnosis cascade

We compose `bibtex-updater` (Stage 1) with each of Claude Sonnet 4.6, GPT-5.1 and GPT-5.4 (Stage 2; up to five tool calls per entry) into a cascade. Stage 1 emits VERIFIED, a typed HALLUCINATED verdict, or UNCERTAIN; only the UNCERTAIN entries are forwarded to Stage 2. Since all cascade results use `bibtex-updater`, the co-design caveat applies (§ G.2). If we treat the entries that neither of the two stages resolves as hallucinated, then on `dev_public` the Sonnet 4.6 and GPT-5.4 cascades both reach DR 0.990, at FPR 0.037 and 0.039 (Tab. 1); the `test_public` and `stress_test` results, and the full panel are in § D.1. The two-stage cascade can therefore be considered a constructive resolution of the false-positive inflation behind failure mode (i).

Takeaway. The cascade is the high-recall configuration at a false-positive rate far below the single-stage harnesses': Stage 2 returns a verdict on roughly half of Stage 1's UNCERTAIN bucket (§ D.1), and the residue is either kept as abstentions (*conservative*) or promoted to a flag (*aggressive*), which lifts recall well above standalone `bibtex-updater`'s for two to six points of FPR (Tab. 22). Choose conservative for the lowest FPR; aggressive somewhat improves recall on harder-tier entries, but the increased FPR requires more manual checks.

5 Discussion

Limitations. HALLMARK scores verifiers separately on 14 hallucination types, 6 sub-tests and 3 difficulty tiers, but our analysis has limitations, ordered here by potential influence (full discussion in § I). Because `bibtex-updater` was co-designed with the taxonomy, it and the cascades built on it should be read as an upper bound on performance (§ G.2), and the FPR-driven deployment bottleneck assumes a venue audit limited by reviewer time (§ 4.4). Most hallucinated entries are synthetic perturbations, and the 108 real-world entries are type-skewed and rated by an automated (LLM) proxy (§ H.6). The core splits are English BibTeX from five ML venues, 2021–2023, supplemented by evaluation-only splits with post-cutoff papers, biomedical and non-ML CS venues, and ChatGPT citations on 42 topics (§ A.3). The post-cutoff resistance of the two latest-cutoff models is descriptive, since recency, provider and contamination are confounded (§ F.4), and hosted-LLM rows are dated snapshots, susceptible to endpoint drift (Tab. 29). Extended-thinking variants were only tested on 100 entries, where larger budgets did not improve F1 once outputs stopped being truncated (§ F.5), and HaRC and verify-citations are excluded because rate limits and blocked lookups keep either from returning its own verdicts at benchmark scale (§ B.2.2). Rankings also depend on the definition of a hallucination, which is set outside the benchmark: counting flags on the four types that cite a real work with wrong details as false accusations puts `bibtex-updater` ahead of the cascade (§ H.1).

Conclusion. Despite these limitations, HALLMARK gives citation verifiers a shared test and one protocol that separates three failure modes (Tab. 20), which helps venues and individual authors decide whether to deploy them. Our experiments suggest that careful design is required to avoid brittleness caused by, e.g., post-cutoff papers or agentic lookups. The two-stage cascade (§ 4.5) points to a promising direction; since its first stage is our co-designed tool, testing the design with other first-stage verifiers is left for future work. Finally, HALLMARK raises a foundational question for the research community: what counts as a hallucinated citation?

AI use statement

Generative-AI tools were used in this work as follows. Among the tasks the ICLR policy requires us to disclose, LLMs generated part of the synthetic data set, namely a subset of the hallucinated entries (controlled plausible fabrications and chimeric titles) generated with GPT-5.1, plus ten `plausible_fabrication` entries generated with GPT-4o-mini, described in § 3.2 (prompts in § B.2.3), and served as the three automated raters of the reliability proxy in § H.6. Anthropic

Claude models—the Opus, Sonnet and Fable families—were used as a coding assistant throughout the implementation of the evaluation harness, the baseline wrappers and `bibtex-updater`, and assisted in interpreting results and in auditing the paper's claims and reported numbers against the released artifacts. No LLM formulated the research questions. The taxonomy and the evaluation protocol were conceived by the authors, including the held-out and stress-test splits; LLMs assisted in instantiating them, turning the authors' taxonomy into concrete type definitions and implementing the protocol. Among the tasks for which disclosure is recommended, the same models assisted with drafting and editing paper sections, with figure-generation code, with literature search, and with reference formatting. Every LLM output that entered the paper, the data, or the code was reviewed by the authors: LLM-generated entries were filtered against bibliographic databases and corrected by the systematic ground-truth audit of § A.3, and reported numbers were recomputed from the released per-entry artifacts. The authors take full responsibility for the content of this submission.

ETHICS STATEMENT

HALLMARK supports scientific integrity by enabling systematic comparison of citation-verification tools, informing venue deployment as LLM-assisted writing becomes standard. A taxonomy of hallucination types could theoretically aid adversaries, but the defensive value of understanding these types outweighs the risk; the taxonomy is derived from publicly documented incidents, not novel attack research. HALLMARK is open-source (MIT), GPU-free, and publicly available at `https://anonymous.4open.science/r/hallmark/` (anonymized for review), ensuring access regardless of institutional resources. The benchmark contains bibliographic metadata only, which is already public: it holds no personal data beyond author names as they appear in published records, and its real-world hallucinated entries come from audits that are themselves public. A flag from any verifier is an accusation against a named author, and at the prevalence of real submissions most flags are false alarms (§ 5); we therefore recommend verifier output as a prompt for human review rather than as a finding, and the released tool routes single-source absence to review instead of flagging it.

REPRODUCIBILITY STATEMENT

Dataset construction, the hallucination taxonomy, and the split definitions are described in § 3.2 and § A. The baseline roster, prompts, model IDs, training cutoffs, decoding parameters, and the agentic tool suite are itemized in § 4.2, § B.2, and § B.2.4; the metric definitions are in § 4.1 and § B.1, and the bootstrap and power-analysis protocol in § B.1. The benchmark and the evaluation harness are released under the MIT license at `https://anonymous.4open.science/r/hallmark/` (anonymized for review), together with per-tool aggregate results. Per-entry predictions are released for the ablations, for every zero-shot LLM row of the main table, and for `bibtex-updater`; DOI-only is summary-only, so its aggregates are re-derivable only by re-running the tool. Three of the 42 released result files carry the SHA-256 of the split they scored and a run timestamp; the rest predate that convention and none records a tool version, so a freshness check reports them as unverifiable rather than as fresh. The `coverage` field in the released aggregates is unreliable—14 of them record 1.0 beside a non-zero `num_uncertain` (worst: `llm_openrouter_deepseek_r1_test_public`, 180 UNCERTAIN, true coverage 0.783)—so the Cov. column of Tab. 1 is read from `results/ablations/coverage_reporting.json` instead (for `bibtex-updater`, from `coverage.json` beside its rescored per-entry verdicts). Most model slugs are floating aliases rather than dated-pinned IDs and the third-party inference endpoints have drifted since collection, so the LLM runs reproduce as dated snapshots rather than as absolute re-runs; § 5 and § B.1 state which results this affects.

REFERENCES

Diletta Abbonato. CheckIfExist: Detecting citation hallucinations in the era of AI-generated content. arXiv preprint arXiv:2602.15871, 2026.

ACL. ACL statement on desk rejecting papers with hallucinated references. ACL 2026, `https://2026.aclweb.org/acl_statement/`, 2026.

Ayush Agrawal, Mirac Suzgun, Lester Mackey, and Adam Kalai. Do language models know when they're hallucinating references? In *EACL*, pp. 912–928. Association for Computational

Linguistics, 2024. doi: 10.18653/v1/2024.findings-eacl.62. URL https://aclanthology.org/2024.findings-eacl.62.

Lakshya A Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J Ryan, Meng Jiang, Christopher Potts, Koushik Sen, Alexandros G. Dimakis, Ion Stoica, Dan Klein, Matei Zaharia, and Omar Khattab. Gepa: Reflective prompt evolution can outperform reinforcement learning, 2026. URL https://arxiv.org/abs/2507.19457.

Hussam Alkaissi and Samy I. McFarlane. Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus*, 15(2), 2023. doi: 10.7759/cureus.35179. URL https://doi.org/10.7759/cureus.35179.

Anonymous. bibtex-updater: Automated BibTeX verification and updating, 2025. URL https://anonymous.4open.science/r/bibtexupdater/. Anonymized for review; released with the anonymized artifact.

Samar Ansari. Compound deception in elite peer review: A failure mode taxonomy of 100 fabricated citations at NeurIPS 2025. arXiv preprint arXiv:2602.05930, 2026.

Amanda Bienz, Carl Pearson, and Simon Garcia de Gonzalo. The case of the mysterious citations. *Computing in Science & Engineering*, 2026. doi: 10.1109/mcse.2026.3713088. URL https://doi.org/10.1109/mcse.2026.3713088.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

dblp Team. dblp computer science bibliography — monthly snapshot XML release of october 2025, 2025. URL https://dblp.org.

DeepSeek-AI et al. DeepSeek-V3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638, 2015. doi: 10.1126/science.aaa9375. URL https://doi.org/10.1126/science.aaa9375.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, 1994.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010. URL https://www.jmlr.org/papers/v11/el-yaniv10a.html.

Dongyang Fan, Sebastien Delsad, Nicolas Flammarion, and Maksym Andriushchenko. HalluHard: A hard multi-turn hallucination benchmark. arXiv:2602.01031, 2026. URL https://arxiv.org/abs/2602.01031.

Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76 (5):378–382, 1971. doi: 10.1037/h0031619.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. RARR: Researching and revising what language models say, using language models. In *Annual Meeting of the Association for Computational Linguistics*, pp. 16477–16508. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.910. URL https://doi.org/10.18653/v1/2023.acl-long.910.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92, 2021. doi: 10.1145/3458723. URL https://doi.org/10.1145/3458723.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pp. 4878–4887, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

Adhiraj Ghosh, Sebastian Dziadzio, Ameya Prabhu, Vishaal Udandarao, Samuel Albanie, and Matthias Bethge. Onebench to test them all: Sample-level benchmarking over open-ended capabilities. In *ACL*, pp. 32445–32481. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1560. URL https://aclanthology.org/2025.acl-long.1560/.

Google DeepMind. Gemini 2.5: Our most intelligent AI model, 2025. URL https://deepmind.google/technologies/gemini/.

HaRC Contributors. HaRC: Hallucinated reference checker, 2024. URL https://pypi.org/project/harcx/.

Moritz Hardt. *The Emerging Science of Machine Learning Benchmarks*. Princeton University Press, 2026. Forthcoming; manuscript available at https://mlbenchmarks.org/.

Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. The dataset nutrition label. In *Data Protection and Privacy*, pp. 1–25. Hart Publishing, 2020. doi: 10.5040/9781509932771.ch-001. URL https://doi.org/10.5040/9781509932771.ch-001.

Xiangkun Hu, Dongyu Ru, Lin Qiu, Qipeng Guo, Tianhang Zhang, Yang Xu, Yun Luo, Pengfei Liu, Yue Zhang, and Zheng Zhang. RefChecker: Reference-based fine-grained hallucination checker and benchmark for large language models. *arXiv preprint arXiv:2405.14486*, 2024. URL https://github.com/amazon-science/RefChecker.

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. *ACM Transactions on Information Systems*, 43(2):1–55, 2025. doi: 10.1145/3703155. URL https://doi.org/10.1145/3703155.

ICLR. A retrospective on the ICLR 2026 review process. ICLR Blog, https://blog.iclr.cc/2026/03/31/a-retrospective-on-the-iclr-2026-review-process/, March 2026.

ICML. Area chair instructions, ICML 2026. https://icml.cc/Conferences/2026/AreaChairInstructions, 2026.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. In *ICLR*, 2025. URL https://openreview.net/forum?id=chfJJYC3iL.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023. doi: 10.1145/3571730. URL https://doi.org/10.1145/3571730.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. SWE-bench: Can language models resolve real-world GitHub issues? In *Proceedings of ICLR*, 2024. URL https://openreview.net/forum?id=VTF8yNQM66.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, et al. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4110–4124. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.naacl-main.324. URL https://doi.org/10.18653/v1/2021.naacl-main.324.

J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977. doi: 10.2307/2529310.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6449–6464. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.397. URL https://doi.org/10.18653/v1/2023.emnlp-main.397.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.741. URL https://doi.org/10.18653/v1/2023.emnlp-main.741.

Mistral AI. Large enough: Mistral Large 2, 2024. URL https://mistral.ai/news/mistral-large-2407/.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using Bayesian binning. In *Proceedings of AAAI*, 2015.

OpenAI. GPT-5.1 Instant and GPT-5.1 Thinking system card addendum, 2025. URL https://openai.com/index/gpt-5-system-card-addendum-gpt-5-1/.

Ben Rank, Hardik Bhatnagar, Ameya Prabhu, Shira Eisenberg, Karina Nguyen, Matthias Bethge, and Maksym Andriushchenko. PostTrainBench: Can LLM agents automate LLM post-training? *arXiv preprint arXiv:2603.08640*, 2026. URL https://arxiv.org/abs/2603.08640.

Delip Rao and Chris Callison-Burch. BibTeX citation hallucinations in scientific publishing agents: Evaluation and mitigation. arXiv preprint arXiv:2604.03159, 2026.

Delip Rao, Eric Wong, and Chris Callison-Burch. Detecting and correcting reference hallucinations in commercial LLMs and deep research agents. arXiv preprint arXiv:2604.03173, 2026.

Abhilasha Ravichander, Shrusti Ghela, David Wadden, and Yejin Choi. HALoGEN: Fantastic LLM hallucinations and where to find them. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1402–1425. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.71. URL https://doi.org/10.18653/v1/2025.acl-long.71.

Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. HalluCitation matters: Revealing the impact of hallucinated references with 300 hallucinated papers in ACL conferences. *arXiv preprint arXiv:2601.18724*, 2026a.

Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. HalluCiteChecker: A lightweight toolkit for hallucinated citation detection and verification in the era of AI scientists. arXiv preprint arXiv:2604.26835, 2026b.

Kaiwen Shi, Weixiang Sun, 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. *arXiv preprint arXiv:2602.23452*, 2026. URL https://arxiv.org/abs/2602.23452.

Nazar Shmatko, Alex Adam, and Paul Esau. GPTZero finds 100 new hallucinations in NeurIPS 2025 accepted papers, 2026. URL https://gptzero.me/news/neurips/. GPTZero analysis of 4,841 accepted NeurIPS 2025 papers, published January 21, 2026.

Maxim Topaz, Nir Roguin, Pallavi Gupta, Zhihong Zhang, and Laura-Maria Peltonen. Fabricated citations: an audit across 2.5 million biomedical papers. *The Lancet*, 407(10541):1779–1781, 2026. doi: 10.1016/S0140-6736(26)00603-3. Erratum in: The Lancet 408(10551):218, 2026, doi:10.1016/S0140-6736(26)01339-5.

verify-citations Contributors. verify-citations: Automated citation verification tool, 2025. URL https://pypi.org/project/verify-citations/.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In *Conference on Empirical Methods in Natural Language Processing*, pp. 7534–7550. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.609. URL https://doi.org/10.18653/v1/2020.emnlp-main.609.

William H. Walters and Esther Isabelle Wilder. Fabrication and errors in the bibliographic citations generated by ChatGPT. *Scientific Reports*, 13:14045, 2023. doi: 10.1038/s41598-023-41032-5.

Zuyao Xu, Yuqi Qiu, Lu Sun, Fasheng Miao, Fubin Wu, Xiang Li, Xinyi Wang, Haozhe Lu, Zhengze Zhang, Yuxin Hu, Jialu Li, Luo Jin, 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. *arXiv preprint arXiv:2602.06718*, 2026.

An Yang et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Appendix

Table of Contents

A Dataset and taxonomy **17**
- A.1 Full taxonomy details . . . 17
- A.2 Real-world incident mapping . . . 18
- A.3 Construction details . . . 19
- A.4 Datasheet for HALLMARK . . . 21

B Evaluation protocol and reproducibility **23**
- B.1 Metrics and statistical analysis . . . 23
- B.2 Evaluation implementation . . . 25
  - B.2.1 Pre-screening layer specification . . . 26
  - B.2.2 HaRC and verify-citations: why neither is in the main table . . . 26
  - B.2.3 Baseline wrappers and prompt template . . . 26
  - B.2.4 Reproducibility: LLM experimental setup . . . 27

C Core results and validity **30**
- C.1 Full `test_public` results . . . 30
- C.2 LLM baseline agreement and calibration . . . 31
- C.3 Small and open-weight models . . . 32
- C.4 Full per-type results . . . 33
- C.5 Validity checks . . . 33
  - C.5.1 Core subset analysis . . . 33
  - C.5.2 Synthetic vs. real-world representativeness . . . 34
  - C.5.3 Multi-defect handling . . . 34
  - C.5.4 Shortcut analysis . . . 35
- C.6 External validity: authentic ChatGPT citations (the Walters–Wilder supplement) 36
- C.7 Comparison and additional figures . . . 38
  - C.7.1 Comparison with related citation efforts . . . 38
  - C.7.2 Additional figures . . . 39
- C.8 Cost and latency . . . 40

D Failure mode (i): agentic aggregation **41**
- D.1 Stage-2 diagnosis cascade: scoring stances and full panel . . . 43

E Failure mode (ii): base-rate precision and deployment **43**
- E.1 Deployment PPV analysis . . . 43
- E.2 Selective prediction and the Coverage column . . . 44
- E.3 Regime-conditional deployment guidance . . . 46

F Failure mode (iii): temporal fragility and calibration **47**
- F.1 Temporal supplement (2024–2025) . . . 48
- F.2 Robustness check: cutoff-aware prompting . . . 50
- F.3 Later-cutoff robustness check: GPT-5.4 on the temporal supplement . . . 52
- F.4 Calibration-vs-memorization controls . . . 53
- F.5 Thinking-budget regime boundary . . . 54

G bibtex-updater: the co-designed reference tool **57**
- G.1 Implementation . . . 57
  - G.1.1 Design goals . . . 57
  - G.1.2 Verification pipeline . . . 57

G.1.3 Deployment modes 57
G.2 Evaluation and co-design bias 58

H Robustness ablations **61**
H.1 What counts as a hallucination changes the ranking 62
H.2 Prompt-sensitivity sweep 63
H.3 GEPA prompt optimization 64
H.4 Threshold and aggregation ablation 65
H.5 Input-format and field leave-one-out 65
H.6 Multi-rater reliability proxy 65

I Extended limitations **67**

J Broader impact and deployment safeguards **69**

K Abbreviations **70**

Appendix roadmap. The appendix follows the paper's three failure modes. § A covers the dataset and taxonomy and § B the evaluation protocol, statistics, and reproducibility setup; § C collects the core cross-split and per-type results together with the validity checks. Each failure mode then has a dedicated home: agentic aggregation (mode i) in § D, base-rate precision and deployment (mode ii) in § E.1, and temporal fragility with its calibration controls (mode iii) in § F. § G documents the co-designed `bibtex-updater`, § H reports the ranking-invariance ablations, and a list of abbreviations closes the appendix.

HALLMARK at a glance — three failure modes of LLM citation verifiers

(i) Agentic aggregation

∼15×

FPR, any-no-match vs. consensus over the same sources (.73 vs. .05)

Five-call harnesses reach DR .97–.99, but at FPR .43–.48: the prompted model flags as soon as a single database returns no match. The lever is a deterministic re-aggregation over three of the harness's four databases; the any-vs-consensus ordering is the finding, not the level. Details in § D.

(ii) Base-rate precision

16.3%

highest independent PPV at a venue-realistic 2% base rate (Gemini 2.5 Pro)

At low prevalence, FPR governs precision (Bayes' rule), not recall: roughly one flag in six from the highest-PPV independent verifier is a true hallucination. Details in § E.1.

(iii) Temporal fragility

.07 vs. .89

post-cutoff FPR on the 448-entry 2024–2025 supplement: Opus 4.7 vs. the cohort maximum

8 of 12 LLMs over-flag 2024–2025 papers ("flag everything unfamiliar"); only the two latest-cutoff models hold (Opus 4.7 .073, Sonnet 4.6 .120). For papers from the past 12 months, no tested tool is reliable without cutoff-aware prompting. Details in § F.

Benchmark 2,526 entries · 14 hallucination types · 3 difficulty tiers · 6 sub-tests per entry (Tab. 5, § 3.2)

Reference `bibtex-updater` v1.2.0, co-designed and excluded from ranking (§ G.2): DR .820 · FPR .051 · F1 .880 · abstains on the ∼13% of entries it cannot back with a record (coverage .87); runs locally for a fraction of a cent per entry, ∼2–3 orders of magnitude cheaper than LLM verifiers (Tab. 1, § E.3).

Deployment regime-conditional guidance: Tab. 26, § E.3.

A DATASET AND TAXONOMY

A.1 FULL TAXONOMY DETAILS

Each hallucination type has a characteristic failure signature (Tab. 2), enabling diagnostic analysis of *why* a tool succeeds or fails.

What is a hallucination? The question seems almost trivial. But looking into the details exposes a lot of nuance: telling what is *not* a hallucination is easy (a bit-by-bit match with a trusted database entry), but the reverse is not. Different spellings, handling of diacritics, name conventions (e.g., how Asian names get represented in Latin alphabets), hyphenation rules, and changing author lists across (preprint) versions complicate the picture and do not afford a clear-cut decision. With these caveats in mind, the main goal of HALLMARK is to analyze potential failure modes without being too pedantic about unconditionally enforcing bit-by-bit matches where humans can also err, even with the best of intentions. This does not mean tool designers should not strive for perfection, but they need to acknowledge these nuances—especially the ones that have a large effect on deployment.

Verification decision tree. The tiers reflect a natural workflow: Tier 1 requires single-field lookups (DOI resolution, venue existence); Tier 2 requires cross-referencing (do authors match the resolved DOI record?); Tier 3 demands semantic reasoning (the title *almost* matches a real paper). A tool that only performs DOI lookups catches Tier 1 but misses Tier 2–3 entirely.

Worked example. An illustrative `near_miss_title` pair (Tier 3)—a real entry and its perturbed counterpart, all other fields unchanged:

```
title = {Attention is All you Need}       % real (Vaswani et al., 2017)
title = {Attention is All you Require}  % near_miss_title
```

The perturbed entry passes the D (DOI), A (authors), V (venue), and F (fields) sub-tests; only T (title exists) and X (cross-DB agreement) fail (Tab. 2), so a DOI-only lookup accepts it, and only title-aware verification catches it.

The taxonomy spans 14 hallucination types across three lookup-difficulty tiers, ordered by the verification effort each type demands. **Tier 1** types are settled by one check, either a local field test (year bounds, placeholder strings) or a single lookup (DOI resolution, venue existence). **Tier 2** types require cross-referencing two or more resolved fields against each other (authors

Table 2: **The HALLMARK hallucination taxonomy: 11 empirically-grounded types plus 3 stress-test types (△) across 3 difficulty tiers.** Each type has a characteristic sub-test failure pattern. Sub-tests: **D**OI resolves, **T**itle exists, **A**uthors match, **V**enue real, **F**ields complete, **X** cross-DB agreement. ✓ = expected pass, ✗ = expected fail, ? = not applicable or varies by entry.

Tier	Type	Description	D	T	A	V	F	X
Easy	fabricated_doi	DOI does not resolve	✗	✓	✓	✓	✓	✗
	nonexistent_venue	Invented conference/journal	?	✓	✓	✗	✓	✗
	placeholder_authors	Generic/fake author names	?	✓	✗	✓	✓	✗
	future_date	Year in the future	?	✓	✓	✓	✓	✗
Medium	chimeric_title	Real authors + fake title	✓	✗	✓	?	✓	✗
	wrong_venue	Correct paper, wrong venue	✓	✓	✓	✗	✓	✗
	swapped_authors	Correct title, wrong authors	✓	✓	✗	✓	✓	✗
	preprint_as_published	arXiv cited as venue paper	✓	✓	✓	✗	✓	✗
	hybrid_fabrication	Real DOI + fake metadata	?	✗	✗	✓	✓	✗
	merged_citation△	Metadata from 2+ papers	?	✓	✗	✗	✓	✗
	partial_author_list†△	Subset of real authors	✓	✓	✗	✓	✓	✗
Hard	near_miss_title	Title off by 1–2 words	✓	✗	✓	✓	✓	✗
	plausible_fabrication	Entirely fabricated, realistic	?	✗	✗	✓	?	✗
	arxiv_version_mismatch△	Wrong version claims	?	✓	✓	✗	✓	✗

†Classified when <50% authors present without "et al." indicator. △Theoretically motivated types; zero real-world instances in current dataset. The schema enum member is `AUTHOR_MISMATCH` (value `swapped_authors`).

Table 3: **Full taxonomy with example BibTeX snippets illustrating each hallucination type.** Red text indicates the hallucinated field.

Tier	Type	Example (hallucinated field in red)
1	fabricated_doi	`doi = {10.9999/nips2024.1847}`
	nonexistent_venue	`booktitle = {Intl. Conf. on Advanced AI Systems}`
	placeholder_authors	`author = {John Doe and Jane Smith}`
	future_date	`year = {2030}`
2	chimeric_title	Real authors, `title = {A Novel Approach...}` (nonexistent)
	wrong_venue	Real paper, `booktitle = {ICML}` (actually NeurIPS)
	swapped_authors	Real title, `author = {Wrong Author List}`
	preprint_as_published	arXiv paper, `booktitle = {NeurIPS}` (never published)
	hybrid_fabrication	Valid DOI resolves, but `title = {...}` doesn't match
3	near_miss_title	`title = {Attention Is All You Want}` (vs. "Need")
	plausible_fabrication	Entirely fabricated, all fields realistic but nonexistent
Stress	merged_citation	Authors from paper A, `title` from paper B, venue from C
	partial_author_list	Real paper, `author = {First and Last}` (middle dropped)
	arxiv_version_mismatch	arXiv preprint cited with wrong venue and shifted year

against the identified paper, venue against the resolved record, preprint against published version). **Tier 3** types require multi-source verification or near-duplicate disambiguation (title near-misses, fully plausible fabrications). Three types (`merged_citation`, `partial_author_list`, `arxiv_version_mismatch`) are theoretically motivated compound failure modes: they keep their difficulty tier—2, 2, and 3 respectively—and appear in every core split, with the `stress_test` split adding depth on them (§ A.3).

A.2 REAL-WORLD INCIDENT MAPPING

We mapped 72 real-world hallucinated citations from three documented incident studies to our taxonomy. Tab. 4 shows the mapping results.

Table 4: **Mapping of 72 real-world hallucinated citations to HALLMARK taxonomy types.** Citations were sourced from GPTZero's NeurIPS 2025 analysis, GhostCite, and HalluCitation.

Taxonomy type	Tier	Count
plausible_fabrication	3	40
fabricated_doi	1	10
chimeric_title	2	8
near_miss_title	3	5
wrong_venue	2	4
swapped_authors	2	3
hybrid_fabrication	2	2
Total mapped		**72**

Most real-world citations map to a single taxonomy type: of the 72 in this table, 57 (79%) map directly, while the remaining 15 exhibit compound failure modes and take their highest-tier applicable type. The benchmark contains 108 real-world entries in total across all splits; this table covers the 72 drawn from three documented incident studies, and the remaining 36 come from additional documented incidents. plausible_fabrication dominates (55%), reflecting the predominant LLM failure mode: models generate coherent but entirely fictional references rather than subtly corrupting real ones. Three taxonomy types—merged_citation, partial_author_list, and arxiv_version_mismatch—have zero real-world instances, which motivates their designation as stress-test types (§ 3.1).

A.3 CONSTRUCTION DETAILS

DBLP scraping. Valid entries were scraped from the DBLP API (dblp.org/search/publ/api) using venue-specific queries for NeurIPS, ICML, ICLR, AAAI, and CVPR. We retrieved BibTeX records, verified DOI resolution via CrossRef, and confirmed title existence in Semantic Scholar. Entries failing any verification step were excluded.

Perturbation pipeline. Systematic perturbations follow deterministic rules per hallucination type:

- fabricated_doi: Replace DOI with a non-resolving DOI using one of 20 non-existent prefixes and four suffix styles (path, identifier, year-indexed, and conference-indexed) to avoid template-detectable patterns.
- nonexistent_venue: Replace venue with an LLM-generated plausible but nonexistent conference name.
- placeholder_authors: Replace author list with common placeholder names.
- future_date: Set year to current year + 5.
- chimeric_title: Keep authors from paper A, replace title with LLM-generated plausible title.
- wrong_venue: Keep all fields but swap venue with a different real venue.
- swapped_authors: Keep title and venue, replace authors with those from a different paper.
- preprint_as_published: Take an arXiv-only paper and add a fabricated venue field.
- hybrid_fabrication: Keep a valid DOI but replace title and authors with fabricated metadata.
- near_miss_title: Modify 1–2 words via six strategies: synonym substitution (POS-safe pairs), plural/singular flipping, British/American spelling swap, abbreviation expansion/contraction (e.g., "RL" ↔ "Reinforcement Learning"), hyphenation toggling (e.g., "self-supervised" ↔ "self supervised"), and article removal.
- plausible_fabrication: Template-based combinatorial generation of a complete, realistic but nonexistent entry (random author combinations, plausible titles, real venues). A separate set of 113 LLM-generated entries (Tab. 15) resembles what an LLM actually emits, but is generated a different way.
- merged_citation: Combine metadata from 2–3 real papers into one entry (e.g., authors from paper A, title from paper B, venue from paper C).
- partial_author_list: Take a real paper and drop one or more middle co-authors, keeping only the first and last.
- arxiv_version_mismatch: Cite an arXiv preprint with a reassigned venue and shifted publication year. Classified as a stress-test type since no real-world instances have been observed.

LLM-generated entries. We prompted GPT-5.1 to generate plausible but fictional citations for types requiring coherent fabrication (plausible_fabrication, chimeric_title,

fabricated_doi). Each prompt requested a BibTeX entry with realistic metadata for a specified ML venue and year range. Generated entries were verified against CrossRef and DBLP: entries with title similarity $\geq 85\%$ (token-sort ratio) and author Jaccard similarity ≥ 0.5 to any real paper were flagged as potential duplicates of real work and excluded. This filtering removed approximately 12% of generated entries, documenting an LLM recall failure rate where the model reproduces real papers rather than fabricating new ones. The remaining entries were assigned sub-test labels based on their hallucination type's expected failure pattern.

Real-world collection. We harvested 72 hallucinated citations from three documented incident studies: GPTZero's NeurIPS 2025 analysis, GhostCite (Xu et al., 2026), and HalluCitation (Sakai et al., 2026a). Each entry was mapped to the closest taxonomy type based on its failure mode (e.g., a citation with a non-resolving DOI was classified as fabricated_doi; a citation to a non-existent paper with plausible metadata was classified as plausible_fabrication). The real-world sample is type-skewed: 55% are plausible_fabrication, reflecting the predominant LLM failure mode in practice. Types without real-world examples (merged_citation, partial_author_list, arxiv_version_mismatch) appear in every core split alongside the 11 empirical types; the stress_test split adds evaluation depth on them (§ 3.1).

Adversarial crafting. We manually constructed entries designed to evade specific detection strategies. These include entries with DOIs that resolve to unrelated papers (hybrid_fabrication), entries combining metadata from multiple real papers (merged_citation), and entries with plausible but non-existent venues chosen to be close to real venue names. Adversarial entries stress-test tool robustness beyond what template-based perturbation achieves.

Quality control. Every generated entry passes through automated validation: (1) BibTeX well-formedness check (valid syntax, and the required fields for the entry's own type: author/title/journal/year for @article, booktitle for @inproceedings, publisher for @book, school for theses, institution for @techreport, and author/title/year for @misc, with journal waived for an @article carrying an arXiv identifier), (2) sub-test label consistency (sub-test ground truth matches the hallucination type's expected failure pattern), (3) cross-validation with the valid entry pool to prevent accidental duplicates.

Extension splits. Four evaluation-only splits are released alongside the main splits, outside the dev/test/hidden partition (Tab. 5). The *temporal probe* (60 entries: 30 valid / 30 hallucinated) is the small pre-supplement probe of post-cutoff behavior. Its valid half pairs 15 papers from 2024 with 15 from 2026, the latter scraped from arXiv ML categories via the arXiv API (scripts/probe_temporal_robustness.py); its hallucinated half (18 from 2024, 12 from 2026) combines 21 perturbation and 9 adversarial entries. The 448-entry temporal supplement (§ F.1) validates the probe's findings at scale; the probe additionally covers 2026 arXiv submissions. The *cross-domain split* (test_crossdomain, 500 entries: 200 valid / 300 hallucinated) probes transfer outside the ML-venue regime the main splits sample: 299 biomedical entries (156 PubMed, 143 bioRxiv) and 201 from non-ML CS venues (OSDI, CCS, CHI, USENIX Security, KDD, SIGGRAPH, and others). Valid entries are scraped; hallucinated entries follow the perturbation pipeline across the taxonomy, including the stress types, with 80/140/80 entries in Tiers 1/2/3; pre-screening uses reference_year 2026 for this split. bibtex-updater's evaluation on this split is in § G.2. The *ChatGPT-citation supplement* (341 entries: 172 valid / 169 hallucinated) converts the hand-coded corpus of Walters & Wilder (2023) into authentic, multidisciplinary ChatGPT hallucinations; its construction and results are in § C.6.
The three sources play complementary roles: perturbation entries are *controlled diagnostic tests* that isolate specific verification capabilities, while LLM-generated entries look like the hallucinations that occur in practice and real-world entries are documented to have occurred; stratifying by generation method lets us assess each independently.

Base-paper overlap across splits. Perturbation entries keep the fields they do not corrupt, so a hallucinated entry can be traced to its base paper by DOI or normalized title. On the public partition the bases are disjoint: 0 of 606 dev_public hallucinated entries match a test_public valid entry, 0 of 519 the other way, the two valid pools share no entry, and two base papers are perturbed in both splits. The auxiliary splits are not disjoint from the public pools: 109 of 244 test_hidden hallucinated entries match a valid entry in dev_public (70) or test_public (39), three test_hidden valid entries duplicate dev_public ones, and 109 of 121 stress_test entries match a valid entry in test_public (64), test_hidden (46), or dev_public (8). Results on

test_hidden and stress_test therefore measure robustness to new perturbations of partly seen base papers, not to unseen papers; the public dev/test comparison is the base-paper-disjoint one.

Table 5: **Dataset statistics by split.** Tier distribution refers to hallucinated entries only. dev_public and test_public cover all 14 hallucination types with $n \geq 30$ for most types (one type on dev_public and six on test_public fall just below, the smallest at $n = 23$); test_hidden covers all 14 types but is too small for an $n \geq 30$ floor; the stress-test split provides additional evaluation depth for three types (the single valid entry is the contamination canary). The bottom block lists the four evaluation-only splits, evaluation sets in their own right that probe regimes the core does not sample: the temporal pair probes post-cutoff behavior on 2024–26 papers (§ 4.4), test_crossdomain covers PubMed/bioRxiv and non-ML CS venues (§ 5), and the ChatGPT-citation supplement carries authentic, hand-verified ChatGPT output spanning the humanities, social sciences, and natural sciences (§ C.6). Each samples a different regime at a different prevalence and is scored separately, so the 2,526-entry total and the dev/test/hidden partition remain the ML-venue core (construction in § A.3).

Split	Valid	Halluc.	Total	Tier 1	Tier 2	Tier 3	Types
dev_public	513	606	1,119	148	279	179	14
test_public	312	519	831	128	238	153	14
stress_test	1	121	122	—	85	36	3
test_hidden	210	244	454	52	115	77	14
Total	1,036	1,490	2,526	328	717	445	14
Extension splits (evaluation-only)							
temporal probe	30	30	60	10	10	10	9
temporal supplement (2024–25)	300	148	448	58	52	38	14
test_crossdomain	200	300	500	80	140	80	14
ChatGPT citations (Walters–Wilder)	172	169	341	—	17	152	4

Quality control and design choices. Every entry undergoes automated validation (field completeness, BibTeX well-formedness, sub-test consistency); see § A.3 for the validation rules and acceptance criteria.

We blind the url field at dispatch, so a verifier never sees it and cannot substitute URL resolution for citation verification. The released corpus still stores the field—114 of the 1,950 public entries and 453 of the 454 test_hidden entries carry one—because blinding was applied when an entry was handed to a tool rather than when it was written. One ad-hoc resume runner read the corpus files directly and skipped that step, so a few checkpoints did ship the field into the prompt. The released prediction records make the exposure measurable: the worst-affected main-table row exposes 7 entries of 1,119, which bounds the shift at under one percentage point of FPR even if every exposed verdict flipped. Blinding is now enforced when an entry is constructed rather than at each dispatch site. We include canary strings for contamination detection (unique fixed tokens embedded in dataset metadata; if a model outputs them verbatim, it has memorized the split). No evaluated model emitted a canary token in any released prediction record. We use uniform type distribution to maximize per-type power; reweighting utilities support prevalence-adjusted evaluation. The API supports generation-method stratification for assessing tool performance on perturbation vs. LLM-generated vs. real-world entries.

What we transfer from HumanEval (Chen et al., 2021) is the unit-test-per-instance structure rather than its execution semantics. HumanEval ships every problem with its own assertion suite, so a failing candidate names the property it violated instead of merely lowering a scalar; HALLMARK ships every entry with its own six declared sub-test outcomes, so a verifier's error localizes to the verification property it missed. The analogy stops at execution: our sub-tests are metadata checks with declared expected values, not code run against hidden inputs, so there is no pass@k analog and no partial credit within an entry. We use uniform type distribution to maximize per-type power; reweighting utilities support prevalence-adjusted evaluation. The API supports generation-method stratification for assessing tool performance on perturbation vs. LLM-generated vs. real-world entries.

A.4 DATASHEET FOR HALLMARK

Following Gebru et al. (2021) and Holland et al. (2020), we provide a datasheet for the HALLMARK dataset.

Motivation. HALLMARK was created to provide a standardized benchmark for evaluating citation hallucination detection tools, motivated by the NeurIPS 2025 incident and subsequent audits.

Composition. The public release contains 2,072 BibTeX entries: 826 valid entries scraped from DBLP and arXiv and 1,246 hallucinated entries generated through perturbation, LLM generation, adversarial crafting, and real-world collection across 14 hallucination types. A held-out `test_hidden` split contains 454 additional entries (210 valid, 244 hallucinated, for 1,036 valid entries in total) (per-entry labels withheld; aggregate label counts in Tab. 5), bringing the grand total to 2,526. Split counts: `dev_public` 1,119 (513 valid / 606 hallucinated), `test_public` 831 (312 / 519), `stress_test` 122 (1 / 121). The dev/test/hidden split is deterministic: stratified by hallucination type and tier with `seed=8042` (recorded in `metadata.json`), so the partition is regenerable from the released corpus. Each entry includes 6 binary sub-test labels. Four evaluation-only splits are released alongside the 2,526-entry benchmark, outside the dev/test/hidden partition: a 60-entry temporal probe, a 448-entry temporal supplement (2024–2025), a 500-entry cross-domain split, and a 341-entry supplement of authentic ChatGPT-generated citations converted from Walters & Wilder (2023) (Tab. 5; the ChatGPT supplement's construction in § C.6).

Collection process. Valid entries were scraped from the DBLP API and verified against CrossRef and Semantic Scholar. Hallucinated entries were generated using the methods described in § 3.2 and § A.3.

Preprocessing/Cleaning/Labeling. BibTeX records were normalized to a consistent field ordering. Unicode characters were preserved. Entries were split into dev/test/hidden sets by stratified sampling across hallucination types and tiers (deterministic seed: see *Composition*). Labels were assigned deterministically by the generation pipeline; valid entries carry ground-truth sub-test labels derived from database cross-checks. The released corpus is post-relabel: a systematic ground-truth audit (`scripts/relabel_ground_truth.py`, commits `32fe9d6` and `1475a2f`) re-resolved entries against live databases and corrected real papers that had been mislabeled `HALLUCINATED`; the per-entry flip log is released at `results/reviewer_experiments/relabel_flips.json`.

Recommended uses. HALLMARK is intended for evaluating and comparing citation verification tools, ablating individual verification sub-tests, and cross-tool ranking via tier-weighted F1 and Plackett-Luce infrastructure.

Non-recommended uses. HALLMARK should not be used to train hallucination generators, to produce convincing fake citations, or for any purpose that violates the MIT license.

Distribution. The dataset is distributed under the MIT license (`https://opensource.org/licenses/MIT`) via GitHub at `https://anonymous.4open.science/r/hallmark/` (anonymized for review). The content release reported throughout this paper is pinned to repository tag `v1.2.0`, which fixes the byte-level split contents—the core splits are byte-identical to the post-relabel `v1.1.1`, and `v1.2.0` adds the evaluation-only extension artifacts—and a `manifest.json` carrying `sha256` digests pins the precomputed baseline-result files, which tag `v1.2.3` moved to `data/v1.2/baseline_results/`. A Croissant 1.0 metadata record is included in the repository root. The `test_hidden` split is not publicly distributed.

Maintenance. The benchmark is maintained by the authors and accepts community contributions through pull requests validated by automated checks for BibTeX well-formedness, sub-test consistency, and duplicate detection. Two version axes are tracked separately. Each entry carries a `schema_version` field (`1.0`) recording the record structure; the schema is stable and the `data/v1.0/` directory name reflects it. The *content* release—which corrections update as point releases—is identified by the `version` field in `metadata.json` (`1.2.0`; the ground-truth relabel landed in `1.1.1`) and the matching repository tag `v1.2.0`. The co-designed reference tool `bibtex-updater` is pinned to git tag `v1.2.0` throughout: both the standalone results (§ G.2) and the cascade Stage 1 (Tab. 22). The cascade Stage 2 is a dated snapshot (Sonnet 4.6 via OpenRouter, 2026-05-31), as LLM endpoints drift (§ 5). The released per-entry `bibtex-updater` verdicts (`results/relabel_delta/btu_v1_2_0_prescreen_fix/bibtexupdater_{dev,test}_public_per_entry.jsonl` alongside `data/v1.0/baseline_results/bibtexupdater_{dev,test}_public.json`) let a reader recompute the `bibtex-updater` aggregates (DR 0.820, FPR 0.051) offline, without re-installing the tool or re-querying external databases. Robustness ablations collected against

third-party inference endpoints carry a snapshot date and an endpoint-drift caveat (§ 5), so they are reproducible as dated within-run deltas rather than as absolute re-runs.

Ethical and legal considerations. HALLMARK contains only bibliographic metadata; it includes no personally identifiable information beyond author names already present in published bibliographic records. No human subjects are involved and no IRB approval is required.

B EVALUATION PROTOCOL AND REPRODUCIBILITY

B.1 METRICS AND STATISTICAL ANALYSIS

The two errors and the summary metrics. The two trade off against different deployment costs. A missed fabrication that reaches print is usually the worse error—costlier than a false alarm a reviewer can dismiss—but which error dominates depends on the regime: prevalence (the fraction of entries that are hallucinated), the cost ratio $c_{\text{FN}}/c_{\text{FP}}$, and reviewer capacity (§ 4.4). We report the **false-negative rate** $= 1 - \text{DR}$ for completeness.

Four further metrics summarize a tool in a single number. **F1-Hallucination** is the harmonic mean of precision and recall on the hallucinated class; it is prevalence-sensitive, so we compare it only within a split. **TW-F1** weights each hallucinated entry by its tier (1/2/3), rewarding the detection of harder hallucinations; the ranking of the three tools we swept (DOI-only, GPT-5.1, and `bibtex-updater`) held across uniform, linear, and quadratic weightings in the sweep we ran, which we now report with a caveat on the sweep itself (§ B.1).

Tier-weighted F1. For entries $\{e_i\}$ with tier weights $w_i \in \{1, 2, 3\}$, predictions $\hat{y}_i$, and labels y_i:

$$\text{TW-Precision} = \frac{\sum_i w_i \cdot \mathbf{1}[\hat{y}_i = y_i = \text{H}]}{\sum_i w_i \cdot \mathbf{1}[\hat{y}_i = \text{H}]}, \qquad \text{TW-Recall} = \frac{\sum_i w_i \cdot \mathbf{1}[\hat{y}_i = y_i = \text{H}]}{\sum_i w_i \cdot \mathbf{1}[y_i = \text{H}]}, \quad (1)$$

where H denotes the hallucinated class, and TW-F1 is their harmonic mean. Valid entries carry no difficulty tier, so false positives contribute uniform weight 1.0 in the TW-Precision denominator. Because false positives carry uniform weight while true positives are tier-weighted, TW-F1 can be inflated by aggressive over-flagging; MCC provides a complementary prevalence-corrected view.

Expected Calibration Error. We partition predictions into $B = 10$ equal-width confidence bins and compute:

$$\text{ECE} = \sum_{b=1}^{B} \frac{|S_b|}{N} \left|\text{acc}(S_b) - \text{conf}(S_b)\right|, \quad (2)$$

where S_b is the set of predictions in bin b, $\text{acc}(S_b)$ is the fraction of correct predictions, and $\text{conf}(S_b)$ is the mean confidence. Adaptive (equal-mass) binning is available in the codebase.

Baseline integration. HALLMARK provides a baseline registry that supports discovery, availability checking, and dispatch for all integrated tools. New baselines register via a decorator pattern, specifying their dependencies and whether they require API keys. A weekly CI workflow re-runs the DOI-only baseline in full and a 50-entry stratified verify-citations sample on the dev split, and validates the pre-computed rate-limited baselines by checksum. The `stress_test` split contains 122 entries: 121 hallucinated plus a single VALID canary entry planted for contamination detection; with one valid entry FPR is not meaningful, so we report only Detection Rate. Read that rate with the split's composition in mind: 75 of the 122 entries (61.5%) are `partial_author_list` or `arxiv_version_mismatch`, both of which describe a real work inaccurately, and every one of the cascade's misses on this split falls in that class.

The statistical procedures below underlie every numerical claim in the paper. Bootstrap CIs accompany the aggregate metrics in Tab. 1 and the per-type breakdown in Tab. 13; paired bootstrap p-values support pairwise tool comparisons in § 4.3; the per-type power analysis (Tab. 6) bounds which type-level differences in Tab. 13 are statistically distinguishable.

Bootstrap confidence intervals.

Scoring conventions for Tab. 1. Small gaps are point-estimate orderings—rankings by the single measured value, with no test that the gap is statistically real: the Sonnet 4.6 / Opus 4.7 F1 gap of 0.3 pp on `dev_public` is not significant (paired bootstrap on the 2026-09-02 per-entry verdicts, 10,000 resamples: 95% CI on the difference $[-1.5, +2.2]$ pp, two-sided p=0.74; § B.1). Abstentions—entries a tool declines to label either way—are handled differently by the two blocks, and the

difference favors the abstaining LLMs. An UNCERTAIN verdict from an LLM verifier is excluded from the DR/FPR/F1 triple, so those rates are computed over the entries the model committed to; `bibtex-updater` resolves its own abstentions to VALID before scoring, so they count against its DR. DeepSeek-R1 shows the size of the gap: 21.7% of its `test_public` verdicts are UNCERTAIN—in its case API error fallbacks rather than deliberate abstentions (Tab. 9)—and excluding them puts its DR at 0.809 where the committed-VALID convention would give 0.644. Coverage and an aggressive re-flagging stance are in § E.2; the Cov. column of Tab. 1 is read from `results/ablations/coverage_reporting.json`, because the released aggregates record `coverage` 1.0 in 14 of 45 files that also report a non-zero `num_uncertain` and so cannot reproduce it. Predictions were collected 2026-05-05, except the Sonnet 4.6 and Opus 4.7 rows of Tab. 1 (2026-09-02, marked §); hosted-LLM rows are dated snapshots subject to endpoint drift (§ 5).

The aggregate metrics in Tab. 1 are accompanied by 95% confidence intervals computed via stratified bootstrap with 10,000 resamples (`seed=42`), wherever a tool's full-coverage per-entry predictions are available. Stratification by hallucination type ensures that the resampled datasets preserve the original type distribution, preventing bootstrap bias from underrepresented types. For reference, GPT-5.1's `dev_public` 95% CIs are: DR [0.812, 0.861]; F1 [0.747, 0.785] (point estimates 0.837 / 0.766). The released result files carry Wilson intervals on the per-type detection rates; their aggregate `*_ci` fields are null, so the bootstrap intervals quoted in this paper are not recoverable from the release and would have to be recomputed from the per-entry dumps. DOI-only is summary-only on `dev_public`—its released aggregates were reconstructed from confusion-matrix counts rather than stored per-entry predictions—so we report its point estimate without a CI or a paired test on that split. `bibtex-updater`'s full-coverage per-entry verdicts are released on both splits and recompute its point estimates exactly; as a co-designed tool it is reported separately from the independent-tool comparison (Tab. 36), so we give its point estimates without a paired bootstrap.

Significance testing. We use paired bootstrap tests (Efron & Tibshirani, 1994) to compare tools pairwise. For each pair of tools (A, B), we resample entries with replacement and compute $\Delta = \text{F1}_A - \text{F1}_B$ for each resample. The two-sided p-value is $\min(1,\, 2 \cdot \min(P(\Delta \le 0),\, P(\Delta \ge 0)))$; taking $2 \cdot P(\Delta \le 0)$ alone returns $p = 1$ whenever the first-named tool is the weaker one. The only headline pair we test is Sonnet 4.6 vs. Opus 4.7. On `dev_public` their F1 gap (+0.003) is not significant (CI $[-0.015, +0.022]$, two-sided p=0.74); on `test_public`: Sonnet's F1 (0.907, CI [0.891, 0.923]) exceeds Opus's (0.889, CI [0.872, 0.906]) by +0.018, one-sided p=0.023, two-sided p=0.045, Cohen's h=0.059: borderline-significant at α=0.05 with a negligible effect size. The bootstrap confidence interval on the difference, $[+0.001, +0.035]$, excludes zero by a margin narrower than the endpoint drift between collections (Tab. 29), and on the earlier collection the sign of this gap is reversed, so we do not read it as establishing an ordering between the two models. The `bibtex-updater`-vs-Sonnet F1 comparison sits across the co-design boundary, so we report it as a point-estimate difference without a paired p-value.

Per-type power analysis. Per-type counts cap the resolution of type-level comparisons. On `dev_public` they run from 26 (`hybrid_fabrication`) to 76 (`plausible_fabrication`) with a median of 41, and on `test_public` from 23 to 66; the smallest buckets are those from which the ground-truth audit reassigned real papers (see Tab. 5 caption). For detection rate, a type with 30 hallucinated entries and 90% true detection rate has a 95% Clopper-Pearson confidence interval of $[0.74, 0.98]$, compared to $[0.55, 0.998]$ at $n = 10$.

Type-level power is modest, so only large per-type gaps are detectable. Tab. 6 reports the minimum detectable effect (Minimum Detectable Effect (MDE)) at 80% power ($\alpha = 0.05$, two-sided z-test) for each hallucination type in the `dev_public` split. A type with $n = 30$ has an MDE of 25.6pp for a single tool; for the difference between two tools it is about 36pp in the worst case (about 22pp for `plausible_fabrication`, $n = 78$), so smaller per-type gaps between tools are not reliably detected. These values bound which per-type comparisons are statistically meaningful.

Tier weight sensitivity. We evaluate tier-weighted F1 under five weighting schemes: uniform $\{1, 1, 1\}$ (equivalent to unweighted F1), linear $\{1, 2, 3\}$ (default), quadratic $\{1, 4, 9\}$, log $\{\log(2), \log(3), \log(4)\}$, and inverse difficulty (weights proportional to $1 -$ mean DR across tools). Tab. 7 reports TW-F1 for the three reference tools across all five schemes.

The `bibtex-updater` column has been re-run on the rescored predictions; the other two have not, and two caveats bound how much they show. First, the implementation that produced Tab. 7 read the false-positive rate under a key the emitter does not write, so it defaulted on every call and the quantity actually swept was tier-weighted *recall*: the sweep could not see a false positive. The lookup is fixed

Table 6: **Per-type minimum detectable effect (MDE) at 80% power on `dev_public`.** Assumes baseline detection rate of 50% (worst-case MDE). Covers the 11 core types (n=495); the three stress-test types (n=111) are omitted.

Tier	Type	n	MDE (pp)
1	fabricated_doi	38	22.7
	nonexistent_venue	39	22.4
	placeholder_authors	41	21.9
	future_date	30	25.6
2	chimeric_title	47	20.4
	wrong_venue	47	20.4
	swapped_authors	67	17.1
	preprint_as_pub.	30	25.6
	hybrid_fabrication	26	27.5
3	near_miss_title	52	19.4
	plausible_fabrication	78	15.9

Table 7: **Tier-weighted F1 under five weighting schemes** (`dev_public`). The `bibtex-updater` column is recomputed as true TW-F1 from the rescored per-entry predictions; its linear value matches Tab. 1. The GPT-5.1 and DOI-only columns are the pre-relabel snapshot (DOI-only's per-entry predictions were not retained, and GPT-5.1's were not re-scored for a post-relabel sweep), whose default-scheme TW-F1 shifts to 0.822 and 0.329 post-relabel (Tab. 1). Their sweep measured tier-weighted *recall* rather than TW-F1 (see text), so only within-column ranges are comparable.

Weighting scheme	DOI-only	GPT-5.1	bibtex-updater
Uniform $\{1, 1, 1\}$	0.392	0.875	0.880
Linear $\{1, 2, 3\}$ (default)	0.310	0.860	0.892
Quadratic $\{1, 4, 9\}$	0.255	0.850	0.901
Log $\{\log 2, \log 3, \log 4\}$	0.337	0.864	0.888
Inverse difficulty	0.342	0.865	0.880
Range (max − min)	0.137	0.025	0.021

in the released harness but the sweep has not been re-run, and no sensitivity artifact is in the release, so read this table as evidence that the *recall* ordering is weight-insensitive and not as evidence about TW-F1 as we define it. Second, the DOI-only column predates the resolver correction of § B.2: its linear cell reads 0.310 where the current run reports TW-F1 0.268 (Tab. 1). Weight-insensitivity is in any case the weaker of the two robustness questions: the ordering does *not* survive a change in what counts as a hallucination (§ H.1). bibtex-updater's narrow post-relabel sensitivity range (0.880–0.901) reflects its consistently strong per-tier performance; DOI-only shows wider variation (0.255–0.392) because its detection is concentrated in Tier 1.

Takeaway. Read the paper's numbers at the resolution the statistics support: per-type cells are too small to separate most pairs of tools (Tab. 6), and the one headline pair with a full paired test—Sonnet 4.6 vs. Opus 4.7—separates by 1.8 pp F1 on `test_public` at a borderline two-sided p=0.045 with a negligible effect size, and not at all on `dev_public` (0.3 pp, p=0.74). Aggregate rankings are the stable quantity within a fixed scoring target: the tool ordering survives all five tier-weighting schemes (Tab. 7) but not a change in what counts as a hallucination (§ H.1), and small per-type and per-pair gaps are point estimates. We report per-type detection rate and not per-type F1 for the same power reason: with 26–76 hallucinated entries per type against 513 valid ones, per-type precision is dominated by the shared false-positive count and tracks type frequency more than detector quality.

B.2 Evaluation implementation

All baselines are implemented as Python wrappers conforming to the HALLMARK baseline interface. Each wrapper: (1) converts HALLMARK `BenchmarkEntry` objects to the tool's expected input format, (2) invokes the tool, (3) maps the tool's output to a HALLMARK `Prediction` with label, confidence, and reason (Fig. 4).

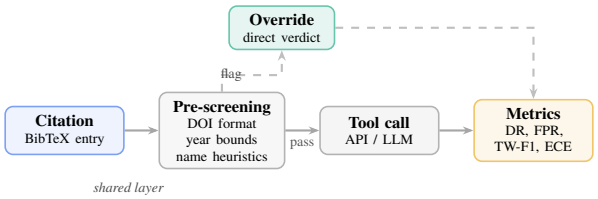

Figure 4: **Evaluation pipeline.** Each citation entry may pass through an optional shared pre-screening layer (DOI format, year bounds, name heuristics) that emits a final verdict directly; otherwise the entry is forwarded to the tool. The layer is applied to the database-lookup tools, whose reported rows therefore include its verdicts: 21 of `bibtex-updater`'s 47 `dev_public` false positives are pre-screening overrides rather than tool verdicts (§ 4.4). The LLM verifiers are run without it, so an FPR is comparable across the two blocks only up to that difference.

B.2.1 Pre-screening layer specification

The pre-screening layer implements four checks that run before external tool invocation:

1. **DOI resolution**: Normalizes the DOI (strips a URL prefix and, for arXiv DOIs, a version suffix such as `v1`, since DataCite registers only the unversioned form) and sends a live HTTP HEAD request to `doi.org`. Only a 404 or 410 served by `doi.org` itself flags the entry; a missing or malformed DOI, a network error, a rate limit, a block or a server error yields no verdict, so a transient failure never overrides the tool.
2. **Year bounds checking**: Flags entries with publication years in the future or before 1900.[1]
3. **Author name heuristics**: Detects common placeholder patterns ("John Doe," "A. Author," single-word author names, repeated names).
4. **Capitalized-token placeholders**: Flags author fields built from placeholder tokens that the previous check misses: a capitalized word followed by digits (`Smith2`), a field made only of initials, or one where more than half the tokens are all-uppercase words.

Pre-screening results are tagged with `[Pre-screening override]` in the reason string to maintain transparency about which detections come from the pre-screening layer vs. the external tool.

B.2.2 HaRC and verify-citations: why neither is in the main table

We attempted HaRC (HaRC Contributors, 2024) and verify-citations (verify-citations Contributors, 2025) as full-coverage baselines and could not obtain a verdict from either at benchmark scale. The binding constraint for HaRC is not Semantic Scholar throttling but `harcx`'s own lookup path, which queries Google Scholar through `scholarly`: Scholar blocks the client, the library retries rather than failing, and under the `-q -threshold 0.75` invocation our wrapper uses even a single-entry `.bib` does not complete within 150 s. Dropping `-q` returns in seconds but emits no verdict line at all, exit 0. An authenticated S2 key (1 RPS, 5-hour budget) does not change either behavior, so 0 of 1,119 entries are actually checked by `harcx`. verify-citations exhibits the same throughput pathology and we never ran it over a full split; the released artifact for it covers a 500-entry sample. In both cases whatever verdicts a run does produce come from the shared pre-screening layer (§ B.2.1) rather than from the tool, so the numbers would not measure what the row claims to measure.

Both tools are therefore excluded from Tab. 1. The one keyed HaRC artifact we release (`harc_with_s2key_dev_public.json`) records DR = 0.209, FPR = 0.045, F1 = 0.335 over all 1,119 entries, but every one of those verdicts is a pre-screening verdict, and the file is scored against the pre-relabel ground truth (633 hallucinated / 486 valid, against the current 606/513), so it is comparable to no other number in this paper and we register it as known-stale rather than report it as a baseline.

B.2.3 Baseline wrappers and prompt template

LLM prompt template. The LLM-based baselines (GPT-5.1, Claude, and OpenRouter models) use a zero-shot prompt with no few-shot examples. The exact template, reproduced verbatim from `hallmark/baselines/llm_verifier.py` (`VERIFICATION_PROMPT`; `{bibtex}` is the substituted entry), is:

[1] Genuinely valid pre-1900 references exist, of course (classical mathematics and physics citations); the bound is a corpus-tuned heuristic—no entry pool in the benchmark carries a publication year before 2021—so deployments verifying older literature should relax or drop this check.

```
You are a citation verification expert. Analyze the following BibTeX
    entry and determine if it is a VALID real publication or a
    HALLUCINATED (fabricated) citation.

BibTeX entry:
```bibtex
{bibtex}
```

Consider:
1. Is the title plausible and does it match known work by these authors?
2. Are the authors real researchers in this field?
3. Is the venue (journal/conference) real?
4. Does the year make sense?
5. If a DOI is present, does it look properly formatted?

When the entry is HALLUCINATED, classify the hallucination mode using
    exactly one of:
`fabricated_doi`, `nonexistent_venue`, `placeholder_authors`, `
    future_date`,
`chimeric_title`, `wrong_venue`, `swapped_authors`, `
    preprint_as_published`,
`hybrid_fabrication`, `near_miss_title`, `plausible_fabrication`,
`merged_citation`, `partial_author_list`, `arxiv_version_mismatch`.
Brief definitions:
- `fabricated_doi`: DOI does not resolve / is invented.
- `nonexistent_venue`: venue/journal does not exist.
- `placeholder_authors`: authors are placeholders ("Author1", "et al."
    alone, etc.).
- `future_date`: year is in the future relative to publication.
- `chimeric_title`: title combines fragments from multiple real works.
- `wrong_venue`: real paper but cited at wrong venue.
- `swapped_authors`: authors swapped or mismatched against the real paper
    .
- `preprint_as_published`: arXiv preprint cited as published in a venue.
- `hybrid_fabrication`: real DOI but other metadata (authors/title) doesn
    't match the DOI target.
- `near_miss_title`: title differs from a real paper by small but
    meaningful edits.
- `plausible_fabrication`: entirely fabricated yet plausible-sounding
    paper.
- `merged_citation`: metadata combined from two real papers.
- `partial_author_list`: real paper but author list is incomplete.
- `arxiv_version_mismatch`: arXiv version cited as a different version (
    or as published).

Respond with JSON only:
{
    "label": "VALID" or "HALLUCINATED" or "UNCERTAIN",
    "confidence": 0.0 to 1.0,
    "predicted_hallucination_type": "<one of the 14 types above, or null
        >",
    "reason": "brief explanation"
}
`predicted_hallucination_type` MUST be null when label is VALID or
    UNCERTAIN.
```

B.2.4 REPRODUCIBILITY: LLM EXPERIMENTAL SETUP

Model roster. The evaluation pipeline accepts any OpenRouter-, OpenAI-, or Anthropic-compatible model ID as a drop-in kwarg (`model=...`). Tab. 8 lists every zero-shot baseline reported in the paper, the API endpoint used for the predictions in `data/v1.0/baseline_results/` and `results/temporal_supplement/`, and the corresponding training cutoff (cross-referenced against Tab. 27). The two Anthropic models are reachable through both the

Table 8: **Zero-shot LLM baseline roster.** Model references: GPT-5.1 (OpenAI, 2025), DeepSeek-V3.2 (DeepSeek-AI et al., 2024), Qwen3 (Yang et al., 2025), Mistral Large (Mistral AI, 2024) and Gemini 2.5 (Google DeepMind, 2025). Endpoint indicates the API used for the released predictions; *native* = vendor SDK, *OR* = OpenRouter mirror. Training cutoffs cross-reference Tab. 27; "≤" marks inferred upper bounds. Run date is the collection date of the released predictions; the two Anthropic rows are a later re-run, collected after the endpoint had drifted (§ 5). `gpt-5.1`, `claude-sonnet-4.6`, and `claude-opus-4.7` are floating aliases (only `gpt-5.4-2026-03-05` is a dated-pinned ID), so every row is reproducible only as a dated snapshot.

Model	Endpoint / model ID	Provider	Cutoff	Run date
GPT-5.1	native: `gpt-5.1`	OpenAI	Sep 2024	2026-05-05
GPT-5.4	native: `gpt-5.4-2026-03-05`	OpenAI	Aug 2025	2026-05-05
Claude Sonnet 4.6	native: `claude-sonnet-4-6`; OR: `anthropic/claude-sonnet-4.6`	Anthropic	≤ Aug 2025	2026-09-02[†]
Claude Opus 4.7	native: `claude-opus-4-7`; OR: `anthropic/claude-opus-4.7`	Anthropic	≤ Oct 2025	2026-09-02[†]
DeepSeek-R1	OR: `deepseek/deepseek-r1`	DeepSeek	Jul 2024	2026-05-05
DeepSeek-V3.2	OR: `deepseek/deepseek-v3.2`	DeepSeek	< Oct 2024	2026-05-05
Qwen3-235B-A22B-2507	OR: `qwen/qwen3-235b-a22b-2507`	Alibaba	Jun 2025	2026-05-05
Qwen3-VL-235B-A22B-Instruct	OR: `qwen/qwen3-vl-235b-a22b-instruct`	Alibaba	≤ Jun 2025	2026-05-05
Mistral Large 2512	OR: `mistralai/mistral-large-2512`	Mistral	≤ mid-2025	2026-05-05
Gemini 2.5 Flash	OR: `google/gemini-2.5-flash`	Google	Jan 2025	2026-05-05
Gemini 2.5 Pro	OR: `google/gemini-2.5-pro`	Google	Jan 2025	2026-05-05
Llama 4 Maverick	OR: `meta-llama/llama-4-maverick`	Meta	≤ Aug 2024	2026-05-05

[†] Re-run on 2026-09-02, later than the other rows, because the OpenRouter Anthropic endpoint drifted after the first run. The numbers can be recomputed from the released per-entry predictions; re-querying the endpoint may not reproduce them (§ 5, § B.1).

native API (baselines `llm_anthropic` and `llm_anthropic_opus_4_7`) and the OpenRouter mirror; we use the OpenRouter mirror for the released predictions. Agentic baselines (`llm_agentic_openai`, `llm_agentic_btu_openai`) use `gpt-5.1`; the reported Anthropic agentic-`bibtex-updater` run (`llm_agentic_btu_sonnet_4_6`) uses the OpenRouter mirror `anthropic/claude-sonnet-4.6`, matching the zero-shot Sonnet endpoint. The registry also retains legacy native-API agentic variants (`llm_agentic_anthropic`, `llm_agentic_btu_anthropic`) hard-coded to `claude-sonnet-4-5-20250929`; these are not used for any number reported in the paper. OpenRouter is accessed via the OpenAI SDK with `base_url=https://openrouter.ai/api/v1`. All zero-shot, DOI-only, and `bibtex-updater` main-run predictions in `data/v1.0/baseline_results/` were collected on 2026-05-05 via the endpoints above; hosted-LLM rows are dated snapshots subject to endpoint drift (§ 5).

Decoding configuration. Zero-shot calls request `temperature=0.0` on every endpoint that accepts it. The one forced exception is `gpt-5.5`, which the code sets to `temperature=1.0`; `gpt-5.1` and `gpt-5.4` are sent `temperature=0.0`, but the GPT-5 endpoint samples stochastically regardless of the requested value (E3 in § F.4 measures the resulting run-to-run variance). The completion-token budget is `max_completion_tokens=1024` on OpenAI- and OpenRouter-compatible endpoints—sized to absorb hidden thinking tokens emitted by reasoning models—and `max_tokens=256` on the Anthropic native endpoint. This single 1024-(non-Anthropic)/256-(Anthropic-native) split applies to every main-table, temporal-supplement, and cutoff-aware run; the thinking-budget smoke test (§ F.5) deliberately sweeps around this budget to locate the saturation boundary. We pass `seed=42` on OpenAI-SDK-compatible endpoints; the Anthropic Messages API does not expose an equivalent. OpenRouter is left on its default dynamic provider routing (no `provider` pin), so an open-weight slug may be served by different underlying providers (e.g. DeepInfra, Together, Fireworks) on different runs; the OpenRouter rows are therefore reproducible only as dated snapshots. The zero-shot prompt (§ B.2.3) is sent as a single user-role message with no system prompt. We do not use JSON mode; responses are parsed with a tolerant extractor that handles bare JSON and fenced code blocks, with an `UNCERTAIN@0.5` fallback on parse failure (throughout, `label@`c denotes the label assigned with confidence c). Agentic calls use a 1024-token completion budget to accommodate tool-call turns. Client-side retry is `max_retries=5` for zero-shot and `3` for agentic baselines, with a 120 s per-request timeout.

Agentic baselines: tools and iteration. We evaluate four agentic variants: OpenAI (`gpt-5.1`) and Anthropic (OpenRouter `anthropic/claude-sonnet-4.6`) backends, each with either a multi-source tool suite or `bibtex-updater`-as-tool. The OpenAI agentic and the Anthropic agentic-`bibtex-updater` runs are the variants reported in the paper; numbers come from `data/v1.0/baseline_results/llm_agentic_btu_sonnet_4_6_dev_public.json` and the corresponding OpenAI checkpoint. Tools are exposed via the provider's native function-calling interface with `tool_choice="auto"`: the model may answer without any tool call, and we track the parametric-vs-tool split per prediction. Up to `MAX_TOOL_CALLS=5` invocations per entry, plus one final verdict turn. If the cap is exceeded without a verdict, the entry is assigned `UNCERTAIN@0.5`.

The multi-source suite comprises 4 tools, each returning {authors, title, venue, year, doi} with fields truncated to 500 characters:
- resolve_doi(doi): CrossRef DOI resolution (api.crossref.org/works/{doi}).
- search_crossref(query, limit=5): CrossRef bibliographic search (max 20 results).
- search_openalex(query, limit=5): OpenAlex polite pool (max 25).
- search_arxiv(query, limit=5): arXiv Atom API (max 20).

All HTTP requests use a HALLMARK/1.0 User-Agent and 10–15 s timeouts; tool exceptions are retried with exponential backoff (up to 3 retries, base delay 1 s).
The bibtex-updater-as-tool variant exposes a single tool, verify_with_bibtex_updater(bibtex), that invokes the bibtex-check CLI with --rate-limit 120 --academic-only, returning {status, confidence, mismatched_fields, api_sources, errors}. Subprocess timeout: 180 s.

External-database access. The three database-backed endpoints—DOI-only, standalone bibtex-updater, and the agentic suite—query five external services: CrossRef (api.crossref.org, polite pool, no key), OpenAlex (polite pool, identified by a mailto contact), Semantic Scholar (the graph/v1 paper-lookup API, accessed with an S2_API_KEY), the arXiv Atom API, and DBLP. Semantic Scholar resolves a citation by matching its title, authors, venue, and identifiers against the S2 corpus and returning the canonical record—or none—which is the evidence bibtex-updater weighs to confirm or flag an entry. Semantic Scholar requires an API key for reliable programmatic access: the free key moves requests off the throttled shared public pool onto a dedicated per-key rate limit. The key affects only throughput and reliability: an unauthenticated re-run hits the shared limit (frequent HTTP 429) and resolves the corpus more slowly, but queries the same records and returns the same matches, so it does not change the reported verdicts. The main-run numbers were queried on 2026-05-04/05. Agentic tool results are replayable offline from the released sha256-keyed tool-cache (mechanics under *Tool-call caching* below), so the agentic baselines reproduce without live calls. The DOI-only and standalone-bibtex-updater endpoints have no such cache, so their numbers are dated snapshots of the corresponding databases.

System prompts. The multi-source agentic system prompt:

```
You are a citation verification expert with access to bibliographic
    lookup tools.
Your task: determine whether a BibTeX entry is a VALID real publication
    or a HALLUCINATED (fabricated) citation.

Strategy:
1. Inspect the entry for obvious red flags (fake DOI prefix, future year,
     placeholder authors).
2. Use tools to cross-reference: resolve the DOI, or search by title/
    author.
3. After gathering evidence (or after finding sufficient signal), emit
    your verdict.

When you are ready to give your final answer, output ONLY valid JSON --
    no prose, no markdown fences:
{
    "label": "VALID" or "HALLUCINATED",
    "confidence": 0.0 to 1.0,
    "reason": "concise explanation citing the evidence you found"
}

Do NOT output the JSON until you have used enough tools or determined
    that parametric knowledge is sufficient.
```

Listing 1: System prompt (multi-source)

The bibtex-updater-as-tool agentic system prompt:

```
You are a citation verification expert with access to a specialized tool,
     'verify_with_bibtex_updater', which cross-references a BibTeX entry
    against CrossRef, DBLP, and Semantic Scholar and returns a structured
     verdict.
```

```
Strategy:
1. For almost every entry, call `verify_with_bibtex_updater` once with
    the exact BibTeX string you were given.
2. Interpret the returned `status` field: statuses like `verified`, `
    url_verified`, or `published_version_exists` suggest VALID; statuses
    like `not_found`, `title_mismatch`, `swapped_authors`, `hallucinated
    `, `future_date`, `doi_not_found`, or `venue_mismatch` suggest
    HALLUCINATED.
3. If the tool returns `api_error` or you suspect the tool is wrong (e.g.
     it reports `verified` but the entry still looks suspicious on
    inspection), apply your own judgment -- the tool is evidence, not an
    oracle.
4. If the first call is unambiguous, do NOT call again. Extra calls waste
     budget.

When you are ready to give your final answer, output ONLY valid JSON --
    no prose, no markdown fences:
{
    "label": "VALID" or "HALLUCINATED",
    "confidence": 0.0 to 1.0,
    "reason": "concise explanation referencing the tool status and any
        disagreement"
}
```

Listing 2: System prompt (`bibtex-updater`-as-tool)

Tool-call caching. Agentic tool results are cached in SQLite keyed on `sha256(tool_name::canonical_args_json)`. TTL is disabled by default; cache hits bypass the network. This is a reproducibility artifact, transparent to the LLM.

Failure handling. After three consecutive API errors, the baseline aborts and marks remaining entries as `UNCERTAIN@0.5` with reason tag `[Error fallback]`. These entries count as abstentions in Cov. and contribute zero to detection rate.

C Core results and validity

C.1 Full `test_public` results

Cross-split observations (from the main text).

- **Cross-split FPR shifts vary by model.** Most LLMs gain or hold F1 on `test_public`; FPR shifts range from under 1 pp (GPT-5.4, both Geminis) to about $+8$ pp (the Qwen variants), with Sonnet 4.6 at $+2.6$ pp and Opus 4.7 at $+5.1$ pp. DeepSeek-R1's -30.3 pp is a denominator artifact—180 of its 831 `test_public` verdicts are API error fallbacks recorded as UNCERTAIN—not a precision gain. All shifts sit within bootstrap-CI width at n_{valid}=312.
- **bibtex-updater is cross-split stable.** Its FPR rises only $+2.9$ pp ($0.051 \rightarrow 0.080$) and DR and F1 hold: conservative matching flags so sparingly that `test_public`'s harder valid pool does not inflate its false positives, and its flat risk–coverage curve—the error rate as a function of the fraction of entries it commits to, traced by varying the abstention threshold—confirms the low FPR is structural, arising from conservative matching rather than from abstention (a reading the coverage re-scoring puts in question; § E.2).
- **Both agentic `bibtex-updater` harnesses lose part of their dev-side FPR inflation on `test_public`.** Sonnet+`bibtex-updater` FPR drops -8.8 pp and GPT-5.1+`bibtex-updater` -11.4 pp cross-split, shrinking each harness's dev-side multiplier, but the harness-driven precision cost persists in both.

Tab. 9 reports all tools with both `dev_public` and `test_public` evaluations: the DOI-only baseline, twelve zero-shot LLM verifiers, three agentic harnesses, `bibtex-updater`, and the always-call evidence-injection co-designed variant.

Takeaway. How much a model's FPR shifts across splits does not depend on how aggressively it flags: GPT-5.4 and both Geminis move under 1 pp, Sonnet 4.6 and Opus 4.7 $+2.6$ and $+5.1$ pp, the recall-aggressive open-weight models $+2.6$ to $+8.2$ pp. Sonnet 4.6 now edges past `bibtex-updater`'s F1 on both splits (0.893 vs. 0.880; 0.907 vs. 0.889), at about three times its FPR (0.161 vs. 0.051).

Table 9: **Full `test_public` (831 entries) results vs. `dev_public` (1,119 entries).** Zero-shot LLMs ordered by $|\Delta\text{FPR}|$. DeepSeek-R1 is the lone zero-shot outlier moving in the opposite direction (-30.3 pp): it records 180 API error fallbacks on `test_public` as `UNCERTAIN` (21.7%, vs. 18/1119 on dev), so its test FPR is measured over 651 answered entries against dev's 1,101. Both `bibtex-updater`-agentic harnesses move in the opposite direction (Sonnet+`bibtex-updater` -8.8 pp, GPT-5.1+`bibtex-updater` -11.4 pp), partly reversing the dev-side harness FPR penalty. In the ΔFPR column, blue marks cross-split-stable tools ($|\Delta\text{FPR}| \leq 1$ pp) and red upward drift ($\geq +6.9$ pp); DeepSeek-R1's -30.3 pp is an artifact of that truncated denominator and is left unshaded.

	dev_public			test_public			
Tool	**DR**	**FPR**	**F1**	**DR**	**FPR**	**F1**	**ΔFPR**
Citation-database baseline							
DOI-only	0.218	0.043	0.347	0.191	0.042	0.314	− 0.001
Zero-shot LLMs							
GPT-5.4	0.767	0.228	0.783	0.780	0.224	0.815	− 0.004
Gemini 2.5 Flash	0.500	0.100	0.631	0.505	0.106	0.644	+0.006
Gemini 2.5 Pro	0.476	0.050	0.627	0.458	0.059	0.613	+0.009
Llama 4 Maverick	0.614	0.146	0.707	0.631	0.167	0.729	+0.020
DeepSeek-V3.2	0.911	0.702	0.727	0.911	0.728	0.776	+0.026
Claude Sonnet 4.6	0.916	0.161	0.893	0.922	0.188	0.907	+0.026
Mistral Large	0.716	0.250	0.742	0.688	0.282	0.741	+0.032
Claude Opus 4.7	0.906	0.154	0.890	0.900	0.205	0.889	+0.051
GPT-5.1	0.837	0.411	0.766	0.852	0.481	0.796	+0.069
Qwen3-VL-235B	0.860	0.551	0.740	0.927	0.628	0.804	+0.077
Qwen3-235B	0.860	0.533	0.744	0.909	0.615	0.798	+0.082
DeepSeek-R1	0.896	0.623	0.739	0.809	0.319	0.812	− 0.303
Agentic harnesses (tool-use; up to 5 tool calls per entry)							
GPT-5.1 + CrossRef/OpenAlex/arXiv	0.967	0.478	0.816	0.942	0.558	0.827	+0.080
GPT-5.1 + bibtex-updater (agentic; tool optional)	0.980	0.470	0.824	0.960	0.356	0.883	− 0.114
Sonnet 4.6 + bibtex-updater (agentic; tool optional)	0.990	0.431	0.841	0.990	0.343	0.902	− 0.088
Co-designed (reference upper bound; § G.2)							
bibtex-updater (v1.2.0)	0.820	0.051	0.880	0.838	0.080	0.889	+0.029
GPT-5.1 + bibtex-updater (always-call; output in prompt)	0.843	0.144	0.856	0.855	0.256	0.851	+0.112

C.2 LLM BASELINE AGREEMENT AND CALIBRATION

The aggregate and per-type numbers for every LLM baseline are in Tabs. 1 and 13. Here we add three diagnostics that the released per-entry `dev_public` predictions support for the open-weight pair Qwen3-235B and DeepSeek-V3.2, which characterize how these verifiers fail rather than how often.

Generation-method stratification. Tab. 10 stratifies detection rate by the entry's generation method, probing whether an LLM verifier has an edge on LLM-generated entries.

Table 10: **Detection rate stratified by generation method on `dev_public`.** LLM-generated entries are the hardest source for all three models, the opposite of what a generator-familiarity advantage would predict.

Generation method	n	**Qwen3-235B**	**DeepSeek-V3.2**	**GPT-5.1**
Adversarial	60	0.917	1.000	1.000
Perturbation	410	0.866	0.937	0.846
Real-world	46	0.978	0.935	0.891
LLM-generated	90	0.730	0.722	0.656
All valid (FPR)	513	0.533	0.702	0.411

For hallucinated entries the metric is detection rate; for valid entries it is the false positive rate over all 513 (486 scraped + 27 relabeled). Values are regenerated on the relabeled `dev_public` split; the relabel reclassified 23 recovered real papers out of the LLM-generated hallucinated bucket (n 113→90). The GPT-5.1 column repeats Tab. 15; GPT-5.1 also generated the LLM-generated entries, so its column is the same-generator case.

LLM-generated hallucinations remain the hardest source for both models (DR 0.722–0.730) relative to perturbation-based entries (DR 0.866–0.937), consistent with the GPT-5.1 finding in § B.1. Were verifiers systematically attuned to LLM-generated text, those entries would be easier to catch; the

reverse holds here. The generator is GPT-5.1 (§ A.3), so for these two models the stratification probes cross-model familiarity rather than strict self-recognition; the same-generator case is the GPT-5.1 stratification in Tab. 15. Adversarial entries are near-perfectly detected (0.917–1.000), so current adversarial strategies do not fool LLM verifiers.

Pairwise agreement. Tab. 11 reports Cohen's κ for all 66 pairs of the twelve zero-shot baselines with stored per-entry `dev_public` predictions (`scripts/compute_pairwise_kappa.py`).

Table 11: **Pairwise agreement between LLM baselines on `dev_public`** (Cohen's κ; n=1,119; abstentions scored as committed-VALID, unlike § 4.3, which excludes them). The two Anthropic columns use the later OpenRouter snapshot (2026-05-30; FPR 0.162/0.165), not the native 2026-09-02 run behind their Tab. 1 rows, and carry its drift caveat (§ E.2). Agreement clusters by disposition: the precision pair (Opus–Sonnet, κ=0.75) and the recall-aggressive cohort (R1–V3.2 0.50, R1–Qwen3-235B 0.53) agree internally, while cross-disposition pairs fall as low as κ=0.15 (V3.2–Gemini Pro). Cells are shaded in proportion to κ (darker = higher agreement).

	Opus 4.7	Sonnet 4.6	R1	V3.2	Flash	Mistral	Qwen3	GPT-5.4	GPT-5.1	Llama	Qwen3-VL
Sonnet 4.6	.746										
DeepSeek-R1	.317	.309									
DeepSeek-V3.2	.243	.229	.497								
Gemini 2.5 Flash	.409	.448	.220	.187							
Mistral Large	.480	.513	.240	.298	.541						
Qwen3-235B	.370	.384	.529	.454	.311	.295					
GPT-5.4	.576	.630	.379	.314	.512	.605	.500				
GPT-5.1	.499	.526	.431	.404	.365	.485	.593	.635			
Llama 4 Maverick	.514	.534	.309	.250	.722	.631	.423	.661	.484		
Qwen3-VL-235B	.361	.368	.555	.495	.303	.313	.773	.496	.558	.412	
Gemini 2.5 Pro	.414	.428	.177	.146	.622	.471	.225	.470	.333	.545	.233

The matrix orders itself by disposition: within-cohort κ runs about 1.6 times (mean 0.54 vs. 0.34, splitting the cohorts at FPR 0.3 in Tab. 1) the cross-cohort values, with GPT-5.1 the exception: it agrees about as much with the low-FPR cohort (0.48) as with its own (0.50), so verifier errors are correlated within cohort and ensembling verifiers of the same disposition adds little independent evidence (cf. the voter ensemble in § H.4). The moderate Cohen's κ (0.454) between DeepSeek-V3.2 and Qwen3-235B indicates that despite similar overall detection rates, the two models disagree on a substantial fraction of entries. Of the 1,119 entries where both models have predictions, they agree and are both correct on 620, agree but are both wrong on 274 (predominantly false positives on valid entries), and disagree on 225. Thirty hallucinated entries are missed by *both* models, a "hard core" that may require a different verification strategy such as API-based cross-referencing.

Confidence calibration. A verifier's confidence is useful only if it separates correct verdicts from wrong ones. For Qwen3-235B and DeepSeek-V3.2 it does not: both assign a median confidence of 0.95 whether the verdict is right or wrong, and their mean confidence on correct predictions exceeds that on incorrect ones by only 0.008 and 0.001 respectively, so no confidence threshold can route their unreliable flags to human review. GPT-5.1's confidences retain some discriminative signal (ECE 0.190, Tab. 1), though far from calibrated. For deployment pipelines that triage flags by confidence, the two open-weight models' scores are effectively uninformative.

C.3 SMALL AND OPEN-WEIGHT MODELS

Setup. We evaluate Claude Haiku 4.5 and the Qwen3 dense ladder at 4, 8, 14 and 32B on `dev_public` under the protocol of § B.2.4: the shared zero-shot prompt, temperature 0, seed 42, one call per entry, no tool access. Haiku runs through OpenRouter on the same path as the main-table Anthropic rows. The Qwen3 models run through the HuggingFace Inference Providers router (`Qwen/Qwen3-{4B,8B,14B,32B}`, provider pinned to `featherless-ai`, collected 2026-08), so the ladder varies parameter count with provider, prompt and decoding settings held fixed, which is what makes the comparison interpretable.

Scaling up does not help. Scaling within the ladder does not buy verification. MCC peaks at the smallest model and the 14B model is the worst of the four, flagging 98.5% of the split against the 4B 95.9%; the 32B model recovers nothing. One entry drew an UNCERTAIN verdict, and the median stated confidence is 0.95 at every model. These models are not failing to answer, and they are not hedging: they answer HALLUCINATED on nearly every entry, with high stated confidence, and MCC places all four near zero.

Table 12: **Small and open-weight verifiers on `dev_public`** (same prompt, temperature 0, seed 42, one call per entry; § B.2.4). Frontier reference rows are repeated from Tab. 1.

Model	DR ↑	FPR ↓	F1 ↑	MCC ↑	Median lat.
Claude Haiku 4.5	0.866	0.444	0.773	0.448	—
Qwen3-4B	0.975	0.940	0.704	0.090	2.5 s
Qwen3-8B	0.977	0.949	0.703	0.074	2.3 s
Qwen3-14B	0.990	0.979	0.703	0.047	3.6 s
Qwen3-32B	0.975	0.947	0.702	0.073	5.2 s
Reference (Tab. 1)					
Qwen3-235B	0.860	0.533	0.744	0.358	—
Claude Opus 4.7	0.906	0.154	0.890	0.755	—

C.4 Full per-type results

Fig. 3 visualizes per-type detection rates as a heatmap; Tab. 13 reports the underlying detection rate for the 11 main types and every full-coverage independent baseline. HaRC and verify-citations are excluded because neither produces its own verdicts at benchmark scale (§ B.2.2).

Per-tier analysis. DOI-only detection concentrates in Tier 1; every LLM drops from Tier 1 to Tier 2 and partly recovers on Tier 3, because `plausible_fabrication`, an invented paper an LLM flags as unfamiliar, makes up 44% of Tier 3. The tiers rank the lookup effort a human needs, not difficulty for an LLM (Fig. 6). The recall–precision tradeoff persists at every difficulty: aggressive models reach their high Tier 1 detection by flagging indiscriminately—their FPR is high—rather than by sharper discrimination, and stay far apart from the conservative models on Tier 3. (Tier 3 aggregates fold in the stress-test type `arxiv_version_mismatch`, so per-type sums in Tab. 13 do not average exactly to the per-tier numbers.)

Per-type analysis: statistical power. Per-type counts run from 26 to 78 on `dev_public`, giving a minimum detectable effect of 22–39 pp between two tools (§ B.1), so within-type cell rankings are directional, not significant.

Per-type failure modes. API tools miss venue-level hallucinations (`preprint_as_published`, `wrong_venue`) because bibliographic APIs do not distinguish venue publication from arXiv availability. GPT-5.1 handles `wrong_venue` (85%) better than `preprint_as_published` (74%), and struggles most with `swapped_authors` (45%) and `near_miss_title` (58%). Prepending `bibtex-updater`'s output to GPT-5.1 cuts calibration error sharply (ECE $0.190 \rightarrow 0.078$) at no recall cost: the strongest option when calibration matters more than raw recall (§ G.2).

Tier concentration. Benchmark optimization rewards easy subtasks (Hardt, 2026). On per-tool, per-tier detection rates, API tools concentrate their wins on Tier 1 (Tier 1/3 ratio $> 30\times$, Gini 0.65), while LLMs perform near-uniformly (Gini < 0.05). We therefore report Tier 3 F1 alongside the aggregate (§ C.4 and Fig. 6) to keep the hard regime visible.
DOI-only with pre-screening (§ B.2.1) detects entries in every type except `preprint_as_published`, where the cited preprint's DOI resolves; without pre-screening it detects only `fabricated_doi` entries. GPT-5.1's two weakest types—`swapped_authors` (0.448) and `near_miss_title` (0.577)—require exact bibliographic recall that even large LLMs lack (parenthesized values are per-type detection rates).

Takeaway. Read the per-type heatmap as a failure-mode diagnostic, not a tool ranking: per-cell MDE between two tools runs 22–39 pp, so per-type gaps between tools are reliable only where they are large; `swapped_authors` and `near_miss_title` stand out as hard for most of the cohort.

C.5 Validity checks

C.5.1 Core subset analysis

Scaling the hallucinated pool does not distort the evaluation signals. The pool grew out of 114 seed hallucinations (71 in dev, 43 in test) through generation across all main hallucination types, and to check the scaling we re-evaluated all baselines on the core subset containing only the 114 seed hallucinations alongside the valid entries. This check was performed on an earlier corpus snapshot: 816 hallucinated (453 dev / 363 test) and 720 valid public entries at the time; it predates the final corpus expansion and the ground-truth relabel, and the released corpus counts are in Tab. 5. On that

Table 13: **Per-type detection rate on `dev_public` for full-coverage independent tools.** HaRC and verify-citations are excluded because neither produces its own verdicts at benchmark scale (§ B.2.2). Most main types have ∼30 instances; correcting the ground-truth mislabels leaves one type (`hybrid_fabrication`, n=26) below 30 (Tab. 5 caption). Every column is scored on the released labels and reproduces the `per_type_metrics` block of the corresponding released result file. Column headers abbreviated for space: S4.6 = Sonnet 4.6, O4.7 = Opus 4.7, G51 = GPT-5.1, G54 = GPT-5.4, R1 = DeepSeek-R1, V3 = DeepSeek-V3.2, Q3 = Qwen3-235B, ML = Mistral Large, GF = Gemini 2.5 Flash, L4 = Llama 4 Mav., GP = Gemini 2.5 Pro, QV = Qwen3-VL. The two red-shaded types are the cohort-wide hard types: the median model catches only 52% and 67% of them, against 81% or more on every other type. Numeric cells are shaded in proportion to the detection rate (darker = higher). The three stress-test types (`merged_citation`, `partial_author_list`, `arxiv_version_mismatch`) are omitted from this table; they appear in every core split and in Fig. 3.

Tier	Type	DOI	G51	G54	S4.6	O4.7	R1	V3	Q3	ML	GF	L4	GP	QV
1	`fabricated_doi`	0.895	1.000	0.974	1.000	0.921	1.000	1.000	1.000	0.974	0.943	1.000	0.703	1.000
	`nonexistent_venue`	0.231	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.974	0.513	0.949	0.571	1.000
	`placeholder_authors`	0.268	1.000	0.927	1.000	1.000	1.000	1.000	0.902	0.927	0.463	0.610	0.805	0.927
	`future_date`	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.857	1.000
2	`chimeric_title`	0.213	0.894	0.872	0.979	0.979	0.867	0.979	0.830	0.809	0.500	0.596	0.614	0.915
	`wrong_venue`	0.298	0.851	0.766	0.936	0.915	0.957	0.915	0.915	0.689	0.587	0.617	0.422	0.851
	`swapped_authors`	0.030	0.448	0.433	0.815	0.940	0.657	0.716	0.522	0.348	0.106	0.194	0.106	0.522
	`preprint_as_pub.`	0.000	0.742	0.581	0.933	0.833	0.900	0.839	0.903	0.452	0.419	0.548	0.100	0.806
	`hybrid_fabrication`	0.154	1.000	0.962	1.000	1.000	1.000	1.000	0.885	0.923	0.692	0.692	0.846	0.923
3	`near_miss_title`	0.135	0.577	0.462	0.673	0.673	0.731	0.788	0.673	0.558	0.220	0.269	0.196	0.712
	`plausible_fabrication`	0.051	0.934	0.895	0.987	0.987	0.974	0.961	0.987	0.842	0.640	0.842	0.680	0.974

snapshot, aggregate metrics differ by less than 2% between the core subset and the full pool across all baselines: detection rate changes by at most 0.015, F1 by at most 0.02, and tier-weighted F1 by at most 0.018. Per-type rankings are preserved: no baseline changes rank on any type. We conclude that the scaled entries are consistent with the seed distribution and do not introduce systematic bias.

C.5.2 Synthetic vs. real-world representativeness

We can detect no surface-feature artifact separating synthetic from real hallucinations, but we cannot yet claim the two are equivalent. For nine `dev_public` types with both real and synthetic entries, per-type Kolmogorov–Smirnov tests on title length, author count, and year fail to reject the null for 6/9 ($p > 0.05$); the three divergent types differ mainly in author count (real entries average 2–3 authors, synthetic 3–4). This is *not* evidence of equivalence: failing to reject a null is not accepting it, the real set has only ∼5–20 entries per type (so the KS test is badly underpowered), and these three surface features are exactly the ones our shortcut analysis (§ C.5.4) shows carry little signal (a logistic regression trained on those features beats the majority-class baseline—always predicting the more frequent label—by only 4.6 pp accuracy, below the 5 pp margin we treat as evidence of an exploitable shortcut). The honest reading is narrow: no artifact strong enough to clear the leakage threshold appears, but distributional and semantic equivalence remains *untested* at these sample sizes. That is the central open validity question, which would need an equivalence test on a much larger real set, not a null-hypothesis test that cannot tell "similar" from "insufficient data". Tool detection is comparable across strata (GPT-5.1: 0.66 on LLM-generated vs. 0.85 on perturbations, with Sonnet 4.6 the same direction), consistent with LLM-generated entries being intrinsically harder rather than with self-recognition. This comparability supports perturbations as controlled diagnostics and motivates expanding the 108-entry real-world set.

C.5.3 Multi-defect handling

The comparison in § C.5.2 covers surface features only; defect count is a structural property it does not reach, and one on which the strata differ sharply. We define a *defect* as a per-field subtest recorded `false`; `cross_db_agreement` is excluded because it is `false` for 100% of hallucinated entries. An entry carries one `hallucination_type` (the injected cause) but may break several fields. Across the three publicly released splits, 330 of 1,246 hallucinated entries (27%) are multi-defect, but the rate is highly stratified by provenance: 13% for `perturbation` (118/892) against 79% for real-world citations (63/80). The real-world set totals 108; the 28 entries in `test_hidden` carry no released subtests, so per-entry rates are computed over the 80 public entries.

Table 14: **Detection rate on single- vs multi-defect hallucinated entries.** A *defect* is a failing per-field subtest. FPR is a single number per model: valid entries have no defects, so it cannot be split by defect count. Split: `dev_public` (430 single-defect, 147 multi-defect, 513 valid). The 29 `future_date` entries have no defect, since only cross-database agreement checks the year (Tab. 2), and are excluded.

Model	DR single	DR multi	Δ	FPR	UNC
GPT-5.4 + BTU (cascade)	0.953	0.952	-0.001	0.039	21
Sonnet 4.6	0.877	0.952	+0.076	0.158	20
bibtex-updater	0.747	1.000	+0.253	0.051	0

We report detection by defect count for three systems spanning the DR–FPR range of Fig. 2 (i.e., the strongest zero-shot LLM (Sonnet 4.6), the rule-based tool (`bibtex-updater`), and the hybrid cascade (GPT-5.4 + BTU)) in Tab. 14. UNCERTAIN counts as a miss, so all denominators are equal; the cascade (21 entries, scored in its conservative variant) and Sonnet 4.6 (20) abstain. Single-defect entries are the harder case for Sonnet 4.6 and `bibtex-updater` ($+0.076$ Sonnet 4.6, $+0.253$ `bibtex-updater`).

C.5.4 Shortcut analysis

Metadata features alone barely predict hallucination labels, so the dataset offers no shortcut a model could exploit instead of reading the content. To test this, we train a logistic regression classifier on eight entry-level features—presence of DOI, field count, author count, title length (characters), title word count, year (numeric), BibTeX entry type, and presence of a `booktitle` field—and evaluate via 5-fold cross-validation on the `dev_public` split.

The logistic regression achieves a cross-validated accuracy of 58.8% (majority-class baseline: 54.2%), a margin of 4.6pp. Since this margin is below 5pp, metadata features provide negligible predictive signal beyond class prevalence. This shows that the metadata features themselves do not leak class information; *which* content fields actually drive detection is quantified separately by the field-leave-one-out ablation (§ H.5), where dropping the title raises FPR by up to 35.5 pp.

Among the metadata features, `has_doi` carries the most weight, consistent with the DOI-only baseline's non-zero detection rate, but no single feature pushes cross-validated accuracy meaningfully above the class-prevalence baseline, confirming that the perturbation pipeline does not introduce systematic metadata artifacts.

Detection rate by generation method. To assess whether GPT-5.1 exhibits self-recognition bias on LLM-generated entries, we stratify detection rate by the entry's generation method (Tab. 15). LLM-generated hallucinations are *harder* for GPT-5.1 to detect (DR = 0.656) than perturbation-based entries (DR = 0.846), the opposite of what a self-recognition advantage would predict, though this stratification cannot separate intrinsic difficulty from a recognition effect it might mask. This is expected: LLM-generated entries are written to look like real citations, so they carry none of the programmatic artifacts that perturbation leaves behind. Perturbation-generated entries dominate the hallucinated pool (410 vs. 90 LLM-generated on `dev_public`, Tab. 15) and are detected $\sim$19 pp more easily, so the public leaderboard is weighted toward this easier class. The harder, content-driven probes of construct validity are the LLM-generated subset and the field-leave-one-out ablation (§ H.5); read absolute detection rates with this class mix in mind. Adversarial entries, despite being crafted to evade detection, are detected without error (DR = 1.000), suggesting that current adversarial strategies do not fool LLM verifiers.

Cross-split validation. We evaluate GPT-5.1 on `test_public` (831 entries, 62.5% hallucinated) to verify that findings generalize beyond `dev_public`. Detection rate is consistent: DR = 0.852 (test) vs. 0.837 (dev). Per-tier patterns are preserved: Tier 1 DR = 1.000, Tier 2 DR = 0.754, Tier 3 DR = 0.831 on `dev_public`. FPR is elevated on both splits (dev: 0.411, test: 0.481), reflecting GPT-5.1's tendency to flag valid entries with non-canonical metadata (uncommon venues, name variants). `bibtex-updater`'s FPR is by contrast cross-split stable (dev 0.051 $\rightarrow$ test 0.080), because its abstention policy declines to flag entries it cannot back with a record rather than guessing on `test_public`'s harder valid pool. MCC = 0.397 (test) summarizes all four cells on that split.

Table 15: **GPT-5.1 detection rate on `dev_public` stratified by generation method.** DR is computed over hallucinated entries, FPR over valid entries; a method with no entries of a given polarity shows "—". The ground-truth audit places 4 perturbation-origin and 23 LLM-generated entries in the valid pool, so those two methods carry a (small-n) per-method FPR alongside the scraped pool (n=486, FPR 0.405); the four per-method FPRs reconcile to the headline `dev_public` FPR of 0.411 over all 513 valid entries. The "Adversarial" generation method in the data files covers `hybrid_fabrication` and `plausible_fabrication` entries that use template-based perturbation (real DOI + fabricated metadata, or combinatorial assembly of plausible fields), distinguishing them from "Perturbation" (simple single-field modifications) and "LLM-generated" (entries produced by language models).

Generation method	n (hall.)	n (valid)	DR	FPR
Adversarial	60	—	1.000	—
Perturbation	410	4	0.846	1.000
Real-world	46	—	0.891	—
LLM-generated	90	23	0.656	0.435
Scraped (valid)	—	486	—	0.405

Takeaway. The validity checks hold: scaling the hallucinated pool from its 114 seeds shifts aggregate metrics by at most 2 pp with no rank changes (§ C.5.1), and a metadata-only classifier exceeds the majority baseline by only 4.6 pp—below the 5 pp leakage bar—so metadata offers no shortcut around reading the content (§ C.5.4). The class mix is the caveat to carry: perturbation entries dominate the pool and are $\sim$19 pp easier for GPT-5.1 than the LLM-generated subset (DR 0.846 vs. 0.656), a direction that also runs opposite to a self-recognition advantage.

C.6 External validity: authentic ChatGPT citations (the Walters–Wilder supplement)

The stratification in Tab. 15 rests on HALLMARK's own LLM-generated entries, so it cannot rule out that our generation pipeline leaves fingerprints a verifier could exploit; an external check needs authentic hallucinations we did not construct. We therefore convert the hand-coded citation corpus of Walters & Wilder (2023) into a fourth evaluation-only split, `supplement_chatgpt_citations` (Tab. 5). Walters and Wilder prompted ChatGPT-3.5 and GPT-4 to write 84 short literature reviews across 42 topics spanning the humanities, social sciences, and natural sciences, then hand-verified every one of the 636 resulting citations, reporting fabrication rates of 55% (GPT-3.5) and 18% (GPT-4). The corpus extends the benchmark along the two axes it is thinnest on: the hallucinations are *authentic* LLM output with expert labels rather than perturbations (the `dev_public` pool behind Tab. 15 has 90 LLM-generated hallucinated entries), and the subject matter is multidisciplinary rather than ML, making it the natural companion to the cross-domain split of § 5.

Conversion. We keep the supported publication type—journal articles—because books, chapters, and websites ($\approx$32% of the source corpus) are structurally unresolvable for the database-backed verifiers evaluated here. A field mismatch on a real work is a hallucination and is mapped to the matching taxonomy type rather than left VALID (Tab. 16); when several fields are wrong, the most identity-defining field wins (title $>$ author $>$ venue). Thirty-four real, findable articles whose only substantive error is a wrong non-future year and/or a volume/page slip are excluded because the taxonomy has no matching type; formatting-only deviations (capitalization, initials-vs-names, missing pages) are likewise not treated as hallucinations. The result is 341 entries (172 valid / 169 hallucinated) that pass the same validation pipeline as the core corpus; the converter, its per-case audit trail, and the raw tool outputs are released with the benchmark code.

bibtex-updater results. On the full supplement (coverage 341/341), `bibtex-updater` reaches DR 0.929 (95% CI $[0.905, 0.953]$), FPR 0.076 ($[0.041, 0.116]$), F1 0.926, and tier-weighted F1 0.959: higher recall than on `dev_public` (0.820) at a somewhat higher FPR (0.051) (Tab. 36) on data from disciplines it was never tuned on. The per-type gradient is: `plausible_fabrication` 1.000 (n=139) and `wrong_venue` 1.000 (n=5), against `near_miss_title` 0.769 (n=13) and `swapped_authors` 0.250 (n=12). The tool is in effect a fabrication detector: an invented work is simply not found in any backing database, whereas a corrupted citation of a real work resolves by title and the tool then accepts the record without re-checking authorship. Stratified by generator, GPT-3.5 output is easier (DR 0.990, FPR 0.053) than GPT-4 output (DR 0.841, FPR 0.078): GPT-4 fabricates less and its errors are subtler, mostly author swaps on real papers. Across subject fields, detection is

Table 16: **Mapping the Walters–Wilder source coding onto the HALLMARK taxonomy.** A substantive field error on a real work maps to the corresponding hallucination type; only wrong-year/volume/page-only errors (34 entries) have no matching type and are excluded.

Source coding	HALLMARK type	Tier	n
work itself fabricated	`plausible_fabrication`	3	139
real work, wrong title	`near_miss_title`	3	13
real work, wrong authorship	`swapped_authors`	2	12
real work, wrong/invented journal	`wrong_venue`	2	5
real work, no substantive error	VALID	—	172

stable (0.85–0.96) while FPR is highest in the natural sciences (0.123 vs. 0.000 humanities, 0.056 social sciences).

Converter audit. Because a conversion artifact is indistinguishable from tool behavior in the aggregate numbers, we audited every false positive and every miss case by case. The audit caught one APA-parser bug—dropped final authors on "A, B, & C" references and deleted ellipsis truncation, which `bibtex-updater` reads as silent author-list truncation—that had inflated DR and FPR simultaneously: spurious flags on valid entries and wrong-reason detections on `swapped_authors`. All numbers above use the fixed converter; the released per-case trail documents both scoring passes. The residual errors are genuine tool behavior: the 13 remaining false positives are metadata sensitivity on real works (strict subtitle/punctuation matching, online-first vs. print years), and all 12 misses are author or title corruptions the tool resolves to the real record and abstains on. An aggregate score on a converted external corpus is thus only as trustworthy as a per-case audit of its false positives and misses.

Zero-shot LLM verifiers. We run four zero-shot LLM verifiers spanning the conservative–aggressive spectrum of Tab. 28 on the supplement, under the default prompt and the main-table protocol (§ B.2.4); every cited work predates all four training cutoffs, so the supplement isolates *parametric* verification—the model verifying a citation against what its weights absorbed in pretraining, without retrieval—on non-ML literature with no post-cutoff confound. Tab. 17 reports the results next to each model's `dev_public` baseline. Three patterns emerge. First, **detection of authentic fabrications separates the models most**: GPT-5.1 (0.964) and Sonnet 4.6 (0.961) approach `bibtex-updater`'s 1.000 on `plausible_fabrication`, while Gemini 2.5 Flash recognizes only 0.345 of them, its overall DR falling from 0.500 to 0.290 even as its FPR stays low (0.029). The drop is consistent with thinner parametric coverage of the multidisciplinary literature, but with four models and no direct probe of what each has memorized, we cannot separate missing coverage from a conservative disposition that declines to flag what it cannot confirm. Second, **the FPR ordering from `dev_public` is preserved and widens**: the spread grows from 0.100–0.533 on ML venues to 0.029–0.860 here, with the aggressive Qwen3-235B flagging 86% of valid multidisciplinary papers; its seemingly strong `swapped_authors` cell (0.917) is a byproduct of that indiscriminate flagging. Third, **the subtle-corruption weakness spans the standalone verifier class**: every calibrated standalone verifier scores at or below 0.25 on `swapped_authors` (`bibtex-updater` 0.250, GPT-5.1 0.250, Sonnet 4.6 0.182, Gemini Flash 0.000); the agentic combination below is the one configuration that improves on it. Stratified by generator, GPT-4 output is harder than GPT-3.5 output for every verifier (e.g. on GPT-3.5 vs. GPT-4 output, Sonnet 4.6 DR 0.94 vs. 0.58; `bibtex-updater` 0.99 vs. 0.84): the better generator fabricates less and corrupts more subtly. Sonnet 4.6 is the only zero-shot model to abstain (15 entries) and posts the best LLM F1 (0.896) with an FPR (0.058) below its ML-venue baseline, consistent with its calibrated profile in the temporal analysis (§ F.1).

GPT-5.1 as a `bibtex-updater` dispatcher. The agentic btu-only harness of Tab. 1—GPT-5.1 with `bibtex-updater` as its only tool, free to decide when to call it and whether to trust it—dominates both of its components on the supplement: DR 0.975 (95% CI $[0.952, 0.994]$), FPR 0.045 ($[0.013, 0.079]$), F1 0.967, tier-weighted F1 0.984, committing on 319 of 341 entries. The combination inherits the tool's perfect fabrication detection (1.000) and improves on both weaknesses the components share: `near_miss_title` rises to 1.000 (tool alone 0.769, model alone 0.385) and `swapped_authors` to 0.571 (4 of 7 committed; both components alone 0.250): the dispatcher treats the tool's mismatch statuses as evidence where the standalone wrapper abstains, and routes genuinely ambiguous cases to `UNCERTAIN`. Notably, the harness FPR inflation of failure mode (i) does not reproduce here: the same configuration that posts FPR 0.470 on `dev_public` posts 0.045 on the supplement. The valid half of this corpus consists of real, database-indexed journal articles

Table 17: **Zero-shot LLM verifiers vs. `bibtex-updater` on the Walters–Wilder supplement (N=341: 172 valid / 169 hallucinated).** DR/FPR on `dev_public` (2021–2023 ML venues) alongside the supplement (authentic, multidisciplinary, pre-cutoff for all models). Abstentions are excluded from DR/FPR (§ B.1); Cov. is the committed fraction (Sonnet 4.6 abstains on 15 entries, the agentic harness on 22; all other verifiers commit on all 341). Fab. is the detection rate on `plausible_fabrication` (n=139), Swap. on `swapped_authors` (n=12). The bottom block is tool-involving: the standalone co-designed tool and the agentic btu-only harness of Tab. 1, whose `dev` columns are that table's agentic (tool-optional) row. Green marks supplement FPRs at or below the `dev_public` value for calibrated verifiers; red the aggressive outlier. *Indiscriminate flagging (FPR 0.860), not discrimination.

	DR		FPR		Supplement			
Model	`dev`	suppl.	`dev`	suppl.	**F1**	**Fab.**	**Swap.**	**Cov.**
Gemini 2.5 Flash	0.500	0.290	0.100	0.029	0.439	0.345	0.000	1.00
Claude Sonnet 4.6	0.916	0.865	0.161	0.058	0.896	0.961	0.182	0.96
GPT-5.1	0.837	0.846	0.411	0.285	0.792	0.964	0.250	1.00
Qwen3-235B	0.860	0.976	0.533	0.860	0.685	1.000	0.917*	1.00
bibtex-updater	0.820	0.929	0.051	0.076	0.926	1.000	0.250	1.00
GPT-5.1 + btu (agentic)	0.980	0.975	0.470	0.045	0.967	1.000	0.571	0.94

that the tool verifies cleanly, so the dispatcher rarely faces the ambiguous partial-match output that produces dev-side over-flagging, consistent with § 4.4's reading that mode (i) characterizes harness behavior on borderline metadata, not the dispatch pattern itself. Protocol note: this run's tool calls used authenticated OpenAlex and Semantic Scholar access, which the standalone `bibtex-updater` run above predates; authentication changes service level (rate limits, latency), while the records consulted are identical.

Takeaway. On original, multidisciplinary ChatGPT-generated citations, two of the benchmark's central findings replicate on data we did not construct: subtle corruptions of real works are the hard class for every calibrated *standalone* verifier (`swapped_authors` DR $\leq$0.25), and FPR—the deployment-decisive metric—spans 0.029–0.860 across the LLM cohort, preserving and widening the `dev_public` ordering. Detection of authentic fabrications spans 0.345–0.964 across LLMs (vs. 1.000 for the database-backed tool), a spread consistent with differences in parametric coverage. The compositional exception is the agentic dispatcher: GPT-5.1 orchestrating `bibtex-updater` reaches DR 0.975 / FPR 0.045 / F1 0.967 and lifts `swapped_authors` to 0.571 (0.333 over all 12 entries), without the mode-(i) FPR inflation it shows on ML venues. Ecological validity cuts both ways: synthetic perturbations neither overstate the difficulty of authentic fabrications nor manufacture the subtle-corruption weakness. Caveat: n=12 for `swapped_authors`, and the supplement is articles-only.

C.7 Comparison and additional figures

C.7.1 Comparison with related citation efforts

Topaz et al. (2026) make the cost of that framing explicit: their 500-entry masked validation, in their own words, "estimates precision but not recall," so the reported prevalence is bounded above but not below—what the pipeline missed is unmeasured. HALLMARK reports both directions on a corpus where every entry carries a ground-truth label, which is the precondition for ranking detectors rather than counting incidents.

Tab. 18 contrasts HALLMARK with concurrent citation-audit efforts. Prior work focuses on auditing published papers; HALLMARK is, to our knowledge, among the first *evaluation benchmarks* for citation-verification tools, alongside concurrent work such as CiteAudit (Shi et al., 2026).

Benchmark design principles. HALLMARK synthesizes design principles from established benchmarks: multi-criteria sub-tests from HumanEval (Chen et al., 2021), temporal segmentation from SWE-bench (Jimenez et al., 2024) and LiveCodeBench (Jain et al., 2025), multi-difficulty challenge sets from Dynabench (Kiela et al., 2021), and continuous, expandable sample pools from ONEBench (Ghosh et al., 2025) that aggregate per-entry measurements while resisting contamination and leaderboard rot. We complement these with the evaluation discipline of PostTrainBench (Rank et al., 2026): pin model versions, timestamp every result, and never assume the newest or largest model is the strongest, since test-set contamination tends to scale with capability.

Table 18: **Comparison of HALLMARK with related citation analysis efforts.** CiteAudit (Shi et al., 2026) is a concurrent, human-validated effort that evaluates citation-verification tools (LLMs and commercial tools); "—" marks attributes we could not verify from the available description. [a]HALLMARK's real-world and adversarial entries receive manual review; perturbation and LLM-generated entries are verified algorithmically (§ 3.2).

	Entries	Types	Sub-tests	Tool eval.	Human-valid.	Open
Citation audits						
GhostCite (Xu et al., 2026)	56K papers	—	✗	✗	—	✗
HalluCitation (Sakai et al., 2026a)	∼300 papers	3	✗	✗	—	✓
Mysterious Citations (Bienz et al., 2026)	HPC venues	—	✗	✗	—	✗
Citation verification benchmarks						
CiteAudit (Shi et al., 2026)	—	—	✗	✓	✓	—
HALLMARK (ours)	2,526	14	6	✓	partial[a]	✓

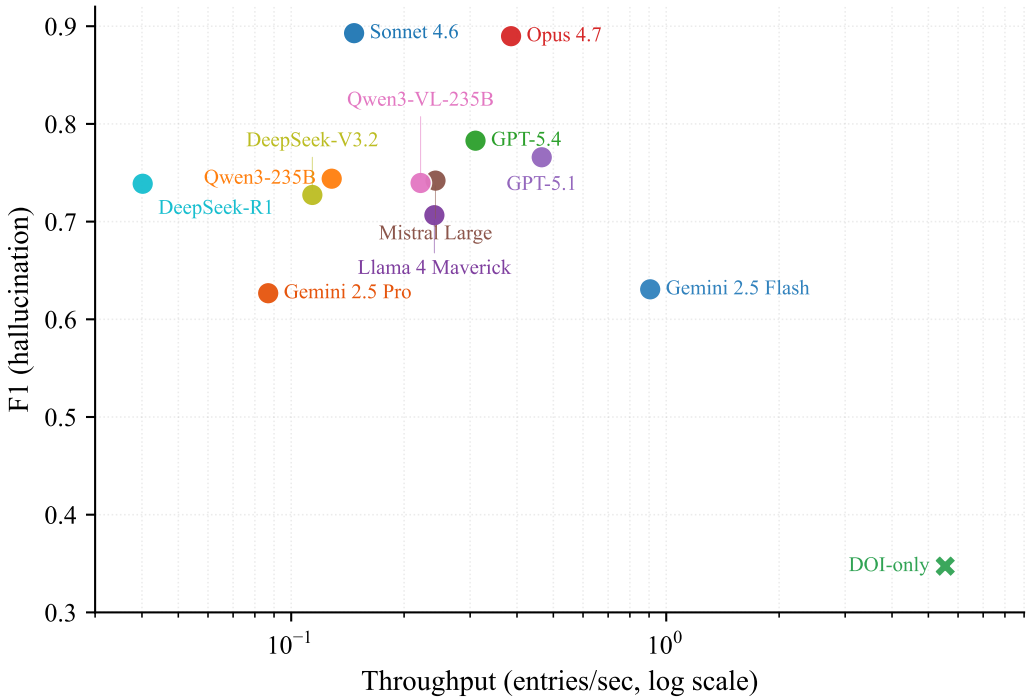

Figure 5: **Accuracy against throughput.** F1 versus measured throughput in entries per second, for the thirteen single-call verifiers that carry timing in the released results. The throughput axis is logarithmic: it spans 0.04 to 5.5 entries/s, and on a linear axis every LLM collapses into the leftmost sixth of the panel. Color distinguishes models only. HaRC and verify-citations are excluded as everywhere else (§ B.2.2); `bibtex-updater` is absent because the release records no timing for it at all (Tab. 19). Throughput is a latency proxy, not a dollar cost; per-entry cost is tabulated separately in Tab. 19. The feasibility constraint at low prevalence (∼2%) is precision rather than either axis: even highly accurate verifiers misallocate reviewer effort because precision is bottlenecked by base rate, not capability.

C.7.2 ADDITIONAL FIGURES

Cost–accuracy tradeoff. LLM verification is also two orders of magnitude more expensive than the rule-based tools, which run locally for a fraction of a cent per entry; single-call LLMs cost cents, and agentic variants add another 2–5× without proportional gains (Fig. 5). With the base-rate PPV drop above, the regime where LLM verification pays off is narrow: high-prevalence settings (triage queues, pre-publication self-check), or pipelines where a downstream human-review stage filters out the false positives.

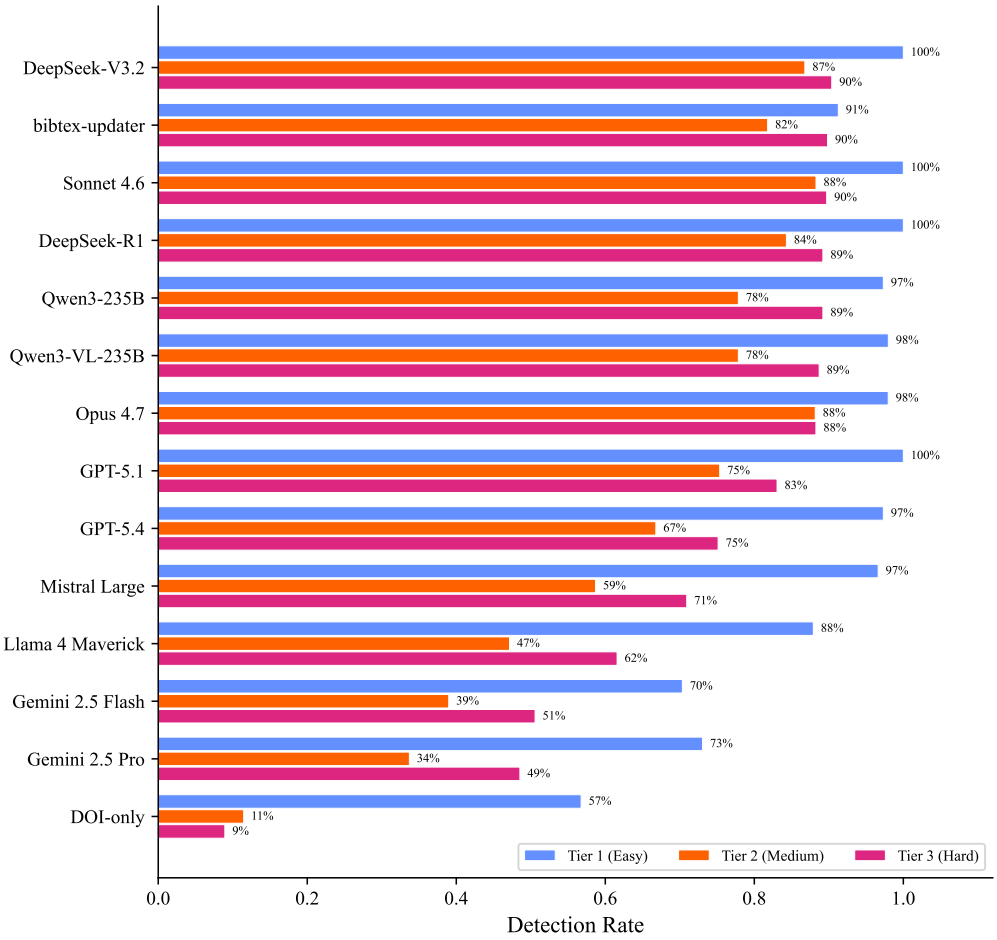

Figure 6: **Detection rate by difficulty tier across all full-coverage tools.** DOI-only concentrates in Tier 1 (.568 against .115 and .089); every LLM drops on Tier 2 and partly recovers on Tier 3.

C.8 Cost and latency

DeepSeek-R1 emits a chain of thought before committing to a verdict while the rest of the zero-shot cohort answers in a single pass, so it is worth stating what the harness holds fixed across the rows of Tab. 1. Four settings are shared by every zero-shot call: the prompt is the verification template of § B.2.3, sent as one user-role message with no system prompt; decoding requests `temperature=0.0`; each call carries `seed=42`; and each call is issued against the same completion budget of `max_completion_tokens=1024` (§ B.2.4).[2] The harness also issues exactly one call per entry, and the measured counts in Tab. 19 bear this out: mean API calls per entry is 1.000 for all twelve zero-shot verifiers, DeepSeek-R1 included, against 2.0–2.9 for the agentic harnesses that may issue up to five lookups. Nothing in the prompt asks for step-by-step reasoning and no reasoning-effort parameter is set for any model in the roster of Tab. 8, so R1's chain of thought is its own decoding behavior inside the shared budget rather than something we request of it. Two limits bound how far this reading goes. The recorded count is of requests the harness issues, so client-side retries inside the SDK are not counted, and those retries are why we read latency off the median rather than the mean. A shared prompt at a shared call budget is also not matched token consumption: R1 spends more of its 1024 tokens on thinking than the single-pass models do, and we ran no matched-compute arm, so Tab. 19 compares each model's deployed configuration at one prompt and one call budget, the setting a user actually faces, and not at equal tokens.

[2]Two configuration caveats, both documented in § B.2.4: the GPT-5 endpoint samples stochastically regardless of the requested temperature (§ F.4 measures the resulting run-to-run variance), and the 256-token budget of the Anthropic native endpoint applies to no released row, since the Sonnet 4.6 and Opus 4.7 predictions were collected through the OpenRouter mirror.

Table 19: **Cost and latency per entry**, recomputed from the released per-entry timing dumps. n is the number of timed entries and identifies the split (1,119 = `dev_public`; 831 = `test_public`, for the five rows reported on `test_public`); latency across splits reflects network and provider conditions, so read down a split, not across. Latency is wall clock around the API call, so it includes client-side retries and provider queueing: the median is the headline, and the p90 tail is mostly backoff. No entry is dropped; the mean, maximum and a 5%-per-tail trimmed mean are in `tables/baseline_cost_latency.csv`. *USD is an estimate, not a measurement*: the zero-shot path records no token counts, so those rows use OpenRouter list prices (2026-05-04) at an assumed 600-prompt/80-completion tokens; agentic rows use their recorded total tokens, split at the same ratio. †Per-entry predictions were not retained, so only the aggregate mean (0.99 s/entry) is recoverable. *No timing was recorded on either split. ‡Lower bound: reasoning tokens are billed but not included in the 80-token assumption.

Verifier	n	Median s	p90 s	API calls	Est. USD/1k
Citation-database tools					
DOI-only†	1,119	—	—	0.434	0.00
Zero-shot LLMs (one API call per entry)					
Gemini 2.5 Pro	831	10.98	20.39	1.000	3.30‡
Gemini 2.5 Flash	1,119	1.06	1.47	1.000	0.38‡
Llama 4 Maverick	831	4.45	19.98	1.000	0.17
Claude Opus 4.7	1,119	2.42	3.42	1.000	5.00
Claude Sonnet 4.6	1,119	7.27	9.18	1.000	3.00
GPT-5.4 (zero-shot)	1,119	2.23	2.95	1.000	1.55
Mistral Large	1,119	3.34	4.20	1.000	1.68
GPT-5.1 (zero-shot)	1,119	2.00	2.73	1.000	1.55
Qwen3-235B	1,119	5.15	17.57	1.000	0.17
Qwen3-VL-235B	831	3.48	6.72	1.000	0.17
DeepSeek-R1	1,119	19.09	29.89	1.000	0.51‡
DeepSeek-V3.2	1,119	7.44	14.78	1.000	0.25
Agentic (tool-use; up to 5 tool calls per entry)					
GPT-5.1 + CrossRef/OpenAlex/arXiv	831	9.22	19.32	2.854	5.61
GPT-5.1 + `bibtex-updater` (agentic)	831	6.25	18.68	2.000	3.42
Sonnet 4.6 + `bibtex-updater` (agentic)	1,119	9.59	14.06	2.059	13.25
Co-designed (reference upper bound; see § G.2)					
`bibtex-updater` (v1.2.0)*	1,119	—	—	1.990	0.00
GPT-5.1 + `bibtex-updater` (always-call)	1,119	2.60	3.96	1.000	1.55

Three readings follow. Chain of thought costs latency and nothing else in the call budget: on `dev_public`, DeepSeek-R1's median of 19.09 s is two and a half times the slowest single-pass model (DeepSeek-V3.2 at 7.44 s) and eighteen times the fastest (Gemini 2.5 Flash at 1.06 s); its $0.51 per thousand entries is a lower bound, since the estimate does not count its reasoning tokens. Latency and dollars therefore rank the cohort differently, and neither ranks it the way FPR does (Tab. 1): the most expensive LLM per entry, Opus 4.7, sits mid-cohort on FPR, while Gemini 2.5 Flash is among the cheapest and has the second-lowest FPR. Finally, wrapping a model in an agentic harness multiplies its estimated cost by 2.2× (GPT-5.1 + `bibtex-updater`) to 4.4× (Sonnet 4.6 + `bibtex-updater`), which buys the recall reported in Tab. 1 at the false-positive inflation analyzed in § D; `bibtex-updater` reaches a comparable operating point with no LLM call at all.

D Failure mode (i): agentic aggregation

A deterministic re-aggregation over three of the harness's four databases confirms the lever: holding sources and matcher fixed, switching from *any-no-match* flagging (flag if any single database returns no match) to *consensus* flagging (flag only when all queried databases agree the entry is absent) cuts FPR from 0.73 to 0.05 (a ∼15× reduction) at a recall cost on metadata-corruption types, which pairing consensus with contradiction checks—`bibtex-updater`'s design—recovers (Tab. 21). So failure mode (i) characterizes this common retrieval-augmented pattern, not every agentic design. Tab. 20 names failure mode (i)—the false-positive inflation of agentic, retrieval-augmented verifiers (Tab. 1). Here we isolate its mechanism: the inflation is a property of the *aggregation rule* that turns multi-database lookups into a verdict, with the base model held fixed, and Tab. 21 quantifies the fix.

Table 20: **Three failure modes that bound LLM-verifier deployment.** Overview results on HALLMARK; with full tables referenced. "1-in-N" is the precision at a venue-realistic $\sim 2\%$ hallucination rate (one true hallucination per N flags).

Failure mode	What breaks	Evidence	Implication
(i) Agentic lookups, diminishing returns (§ 4.3)	The prompted model tends to flag an entry as soon as *any* one database returns no match, so partial database coverage becomes a false positive.	A 5-call budget lifts recall past the conservative rule-based reference (DR .97–.99 vs. .82) but at $\sim 9\times$ its false-positive rate (.43–.48 vs. .05); the rise comes from the harness (any-no-match flagging), not the base model.	More tool calls buy recall, not precision.
(ii) Base-rate precision drop (Tab. 23)	At a venue-realistic $\sim 2\%$ base rate, precision is governed by the ratio of detection rate to FPR (Bayes' rule), and FPR is the term that varies across verifiers.	FPR spans .05–.70 across LLM verifiers, a $\sim 6\times$ precision gap at a 2% base rate: the most precise verifiers reach 1-in-6 (Gemini 2.5 Pro) to 1-in-10 (Sonnet 4.6) flags, high-FPR open-weight models 1-in-35 to 1-in-39.	FPR, not recall, is the deployment-decisive lever.
(iii) Post-cutoff calibration breakdown (Tab. 28)	Most LLMs over-flag papers past their training cutoff: "flag everything unfamiliar".	On 2024–2025 papers, 8 of 12 LLMs degrade sharply (FPR .59–.89), two over-flag only moderately, and only the two latest-cutoff models hold (Sonnet 4.6 .12, Opus 4.7 .07), confounded with possible recall.	Verifier trust expires with the training cutoff.

Cross-source aggregation on the agentic databases. Failure mode (i) attributes the agentic harness's inflated FPR to its emergent *any-no-match* disposition; we isolate that aggregation rule directly. Holding the databases and the matcher fixed, we query three of the harness's four databases—CrossRef, OpenAlex, and arXiv—for every `dev_public` entry, count a source as confirming when it returns a record whose title matches the entry (≥ 0.85 similarity), and score three rules over identical evidence. *Any-no-match* (flag if *any* source misses) reaches FPR 0.729 at DR 0.835; *consensus* (flag only if *all three* miss) cuts FPR to 0.049—roughly $15\times$ lower—at DR 0.251; *majority* (flag if fewer than half confirm) falls between, at FPR 0.201 / DR 0.417. The FPR reduction is therefore a property of the aggregation rule, not the model or the sources. The recall cost is expected and instructive: title-presence consensus catches only *non-existence* hallucinations and misses the metadata-corruption types (`near_miss_title`, `chimeric_title`, `wrong_venue`), where a real paper's title still matches—so the principled rule is consensus-absence *paired with* positive-contradiction checks, exactly `bibtex-updater`'s design (DR 0.820 / FPR 0.051). Two caveats bound the reading: Semantic Scholar, the harness's fourth backend, is excluded because its API key returned HTTP 403 on every endpoint during our evaluation—whereupon the harness silently falls back to a rate-limited keyless pool—so we aggregate over the other three;[3] and the deterministic title matcher is a proxy for the harness's per-source judgment, so the absolute any-no-match FPR need not equal the LLM harness's (and the proxy DR 0.835 likewise understates the harness's 0.97–0.99); the any-vs-consensus *ordering* is what transfers, not the level.

Takeaway. Failure mode (i) is an *aggregation* failure: on identical lookups with the base model held fixed, flagging on any single-source miss gives FPR 0.729, whereas requiring consensus absence lowers it $\sim 15\times$ to 0.049, at a recall cost that pairing consensus with positive-contradiction checks (`bibtex-updater`) recovers (DR 0.820 / FPR 0.051). The figures are a deterministic title-match proxy over three of the harness's four databases, so the any-vs-consensus *ordering* is the finding; the absolute level is proxy-dependent.

[3] This backend fragility is itself an instance of failure mode (i). Semantic Scholar's free API tier has been curtailed since 2024 and its public support repository was archived in January 2025 (`https://github.com/allenai/s2-folks`); separately, OpenAlex began requiring an API key in early 2026, so the harness's keyless OpenAlex path no longer functions either. A verifier that reads such a failed lookup as "no record found" converts backend unavailability into a false positive.

Table 21: **Cross-source aggregation on the agentic harness's databases** (`dev_public`; deterministic title-match proxy over CrossRef/OpenAlex/arXiv at ≥ 0.85 similarity). Holding databases and matcher fixed, the aggregation rule alone moves FPR by $\sim$15×. `bibtex-updater` pairs consensus-absence with positive-contradiction checks, recovering the recall that title-only consensus loses. Read the any-vs-consensus *ordering*, not the absolute level (the proxy differs from the LLM harness's own operating point; Tab. 1). Shading marks the lever: red is the any-no-match FPR and blue the consensus FPR on identical evidence ($\sim$15× apart); the gray row is the co-designed reference, as in Tab. 1.

Aggregation rule	**DR ↑**	**FPR ↓**
Any-no-match (flag if *any* source misses)	0.835	0.729
Majority (flag if fewer than half confirm)	0.417	0.201
Consensus (flag only if *all three* miss)	0.251	0.049
`bibtex-updater` (consensus-absence + contradiction)	0.820	0.051

D.1 Stage-2 diagnosis cascade: scoring stances and full panel

All three cascades share one Stage 1. It decides 882 of the 1,119 `dev_public` entries on its own (384 HALLUCINATED, 498 VALID) and forwards 237 to Stage 2. Sonnet 4.6 returns 151 HALLUCINATED and 2 VALID verdicts and leaves 84 unresolved; GPT-5.1 resolves all 237 (226 HALLUCINATED, 11 VALID), so its conservative and aggressive rows coincide; GPT-5.4 returns 214 HALLUCINATED and 1 VALID and leaves 22 unresolved. Only about a fifth of the split reaches the agentic stage, which is part of why the harness's false-positive inflation does not appear: the routing limits both what tool use can fix and what it can break.

We report two stances: *conservative* keeps the entries Stage 2 still cannot resolve (its residual UNCERTAIN verdicts) as abstentions; *aggressive* flags any entry not affirmatively verified by either stage, assigning it HALLUCINATED at confidence 0.55. With Sonnet 4.6 as Stage 2, aggressive promotion raises Tier-3 F1 by one to two points and pushes recall toward 1.0, at a few points of added FPR and a $\sim$1.5 pp Area Under the ROC Curve (AUROC) drop from collapsing confidences to the fixed 0.55 (Tab. 22). The stance matters only in proportion to the residue Stage 2 leaves: Sonnet leaves 140 of 1,119 `dev_public` entries unresolved, GPT-5.4 leaves 21 (so its two stances differ by at most .003 on any metric), and GPT-5.1 leaves none, which collapses the choice entirely.

E Failure mode (II): base-rate precision and deployment

E.1 Deployment PPV analysis

Published findings pin the base rate only loosely: incident studies report affected-*paper* counts rather than rates (GPTZero: 53 of 4,841 NeurIPS 2025 papers (Shmatko et al., 2026)), and even the implied paper-level rate sits well above the *entry-level* rate the precision calculation needs, so prevalence should be treated as an estimate (for an ablation, see Tab. 24).

At venue-realistic prevalence (1–5%), positive predictive value drops sharply. Tab. 23 reports PPV $= \frac{\text{DR}\cdot\text{prev}}{\text{DR}\cdot\text{prev}+\text{FPR}\cdot(1-\text{prev})}$, where prev is the entry-level hallucination prevalence, for the full-coverage tools and the co-designed `bibtex-updater` (whose DR/FPR use the headline committed-VALID convention; its abstention is reported separately in § E.2).

At a 2% base rate, GPT-5.1 achieves only 4.0% PPV: roughly 24 of every 25 flagged entries are false alarms. Aggressive models (DeepSeek-R1/V3.2) fall below 3% despite $\geq$89% DR. Gemini 2.5 Flash more than doubles GPT-5.1's PPV (9.3% vs. 4.0%) through its much lower FPR, despite a lower detection rate: the precision-bottleneck logic favors the conservative model at deployment-realistic prevalence.

Takeaway. FPR alone does not fix the order: DOI-only has the lowest FPR in the cohort yet lands mid-table at 9.4%, because it catches few hallucinations. The ratio of detection rate to FPR decides.

Prevalence sweep. To separate the *ranking* of tools (which is prevalence-invariant because DR and FPR are) from *deployability* (absolute PPV, which is not), Tab. 24 sweeps the same closed-form

$$\text{PPV}(\text{prev}) = \frac{\text{DR}\cdot\text{prev}}{\text{DR}\cdot\text{prev}+\text{FPR}\cdot(1-\text{prev})} \tag{3}$$

over prevalence values $\{0.005, 0.01, 0.02, 0.05, 0.10, 0.20, 0.30\}$ for four representative tools spanning the FPR spectrum, using the exact DR/FPR from Tab. 1.

Table 22: **Cascade results by Stage-2 backbone and scoring stance.** Rows are grouped by split, then by backbone, then by how residual UNCERTAIN is scored: *cons.* keeps it as an abstention, *agg.* promotes it to a flag at confidence 0.55. All three cascades share one Stage 1; the Sonnet 4.6 and GPT-5.4 Stage 2 runs were collected on 2026-09-21 (Sonnet 4.6 via OpenRouter, a dated snapshot of a drifting endpoint; § 5). "—" marks splits with no scored VALID entries (`stress_test` is evaluated on its 121 hallucinated entries; the split's single valid entry is the contamination canary). `T3-F1` is F1 on Tier-3 (hard) hallucinations.

Split	Stage-2	Mode	DR ↑	FPR ↓	F1 ↑	TW-F1 ↑	T3-F1 ↑	AUROC ↑
`dev_public` (n=1,119)	Sonnet 4.6	cons.	0.989	0.037	0.976	0.982	0.910	0.977
		agg.	0.990	0.037	0.980	0.986	0.935	0.975
	GPT-5.1	cons.	0.974	0.039	0.970	0.976	0.929	0.958
		agg.	0.974	0.039	0.970	0.976	0.929	0.958
	GPT-5.4	cons.	0.990	0.039	0.978	0.985	0.930	0.973
		agg.	0.990	0.039	0.979	0.985	0.933	0.973
`test_public` (n=831)	Sonnet 4.6	cons.	0.989	0.058	0.975	0.982	0.905	0.976
		agg.	0.990	0.058	0.978	0.984	0.928	0.972
	GPT-5.1	cons.	0.965	0.058	0.965	0.971	0.924	0.947
		agg.	0.965	0.058	0.965	0.971	0.924	0.947
	GPT-5.4	cons.	0.990	0.055	0.979	0.985	0.927	0.971
		agg.	0.990	0.058	0.978	0.984	0.928	0.970
`stress_test` (n=121)	Sonnet 4.6	cons.	0.966	—	0.983	0.981	0.971	—
		agg.	0.967	—	0.983	0.982	0.971	—
	GPT-5.1	cons.	0.851	—	0.920	0.927	0.971	—
		agg.	0.851	—	0.920	0.927	0.971	—
	GPT-5.4	cons.	0.967	—	0.983	0.982	0.971	—
		agg.	0.967	—	0.983	0.982	0.971	—

Table 23: **Positive predictive value at venue-realistic prevalence.** In the $PPV_{2\%}$ column, green marks roughly one true catch per six flags (PPV $\geq$16%) and red under one per twenty (PPV $<$5%); the gray row is the co-designed reference.

Tool	DR	FPR	$PPV_{2\%}$	$PPV_{5\%}$
bibtex-updater (v1.2.0)	0.820	0.051	24.8%	46.0%
Gemini 2.5 Pro	0.476	0.050	16.3%	33.4%
Opus 4.7	0.906	0.154	10.7%	23.6%
Sonnet 4.6	0.916	0.161	10.4%	23.0%
DOI-only	0.218	0.043	9.4%	21.1%
Gemini 2.5 Flash	0.500	0.100	9.3%	20.8%
Mistral Large	0.716	0.250	5.5%	13.1%
GPT-5.1	0.837	0.411	4.0%	9.7%
Qwen3-235B	0.860	0.533	3.2%	7.8%
DeepSeek-R1	0.896	0.623	2.9%	7.0%
DeepSeek-V3.2	0.911	0.702	2.6%	6.4%

The sweep makes the deployment lesson concrete: at venue-realistic 1–2% even the best tool flags four-to-five false alarms per true catch, but the *relative* ordering a benchmark establishes transfers unchanged to any deployment prevalence: only the operating point on the PPV curve moves.

E.2 Selective prediction and the Coverage column

Structural precision (from the main text). `bibtex-updater`'s low false-positive rate is *structural*: it comes from conservative matching, not from abstention. It abstains on the ∼13% of entries it cannot back with a record; on the entries it commits to, its FPR is essentially unchanged while its detection rate increases (Tab. 25); abstention raises recall and leaves precision unchanged. Because its low FPR does not depend on abstention, `bibtex-updater` serves as the precision-anchored reference against which the LLMs' abstention-driven gains are measured, and the reason we keep it out of the ranking (§ G.2). Because DR and FPR are prevalence-independent, we compare on them, and on F1 only within a split.

The Coverage column of Tab. 1 reports, for each tool, the fraction of entries on which it *commits* to a VALID/HALLUCINATED verdict rather than abstaining (UNCERTAIN). We report it because

Table 24: **PPV vs. prevalence for four representative tools** (Eq. (3); DR/FPR from Tab. 1). (a) The *ranking* is essentially prevalence-invariant: Opus 4.7 > Sonnet 4.6 > GPT-5.1 > DeepSeek-V3.2 at every prevalence, because PPV is monotone in DR/FPR and the FPR gaps dominate. (b) *Absolute* PPV (deployability) changes sharply with prevalence: the same tool moves from near-useless at venue-realistic prevalence to usable as prevalence rises (Opus 4.7: 2.9% at a 0.5% base rate to 71.6% at 30%). Shading in the 2% column mirrors Tab. 23; at other prevalences the ranking is unchanged and only deployability moves.

Tool	DR	FPR	.5%	1%	2%	5%	10%	20%	30%
Claude Opus 4.7	0.906	0.154	2.9%	5.6%	10.7%	23.6%	39.5%	59.5%	71.6%
Claude Sonnet 4.6	0.916	0.161	2.8%	5.4%	10.4%	23.0%	38.7%	58.7%	70.9%
GPT-5.1	0.837	0.411	1.0%	2.0%	4.0%	9.7%	18.5%	33.7%	46.6%
DeepSeek-V3.2	0.911	0.702	0.6%	1.3%	2.6%	6.4%	12.6%	24.5%	35.7%

abstention is a usable deployment action—defer to a human reviewer—and because it is the lens under which the co-designed `bibtex-updater` reads cleanly: it abstains on the 12–13% of entries it cannot verify against a backing record (coverage 0.87 on dev, 0.88 on test). Its low headline FPR (0.092) does not come from that abstention: on the entries it commits to, its selective FPR (0.099) is essentially unchanged from full coverage, so abstention raises recall and leaves precision unchanged. Where the LLM verifiers lower their false-positive rate as they abstain—precision-via-abstention—`bibtex-updater`'s risk–coverage curve is flat: its precision is *structural*—it comes from conservative matching, not from abstention (its selective FPR 0.099 ≈ its full-coverage 0.092)—so selective prediction adds little. We use "structural" throughout in this operational sense; it does not mean invariant, as the FPR still rises modestly across splits (0.092 → 0.115; Tab. 9). No peer citation benchmark reports coverage; the selective-prediction framing is standard elsewhere (El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017).

Which entries count as abstentions. For the LLM verifiers an abstention is an UNCERTAIN verdict. For `bibtex-updater`, an entry counts as an abstention when its status leaves the tool undecided (`unconfirmed`, or a failed or skipped lookup), or when the tool returned no record for it. The tool's own `abstained` flag is not used, since it also marks `not_found` detections. This gives coverage 0.86 on `dev_public` and 0.88 on `test_public`. The `coverage` field in the released aggregates is unreliable in 14 of 42 files (§ B.1), so the Cov. column of Tab. 1 is read from `coverage_reporting.json` (for `bibtex-updater`, from its released aggregate) rather than recomputed from the release.

Dual scoring. We score each tool three ways (Tab. 25). The *selective* stance scores only the committed (non-abstained) entries; the *conservative* stance scores the full split with every abstention forced to VALID; the *aggressive* stance scores the full split with every abstention forced to a flag (`HALLUCINATED@0.55`, i.e. scored as a flag with confidence 0.55, just above the 0.5 decision threshold). Tab. 1 and Tab. 9 report each tool's native stance (selective for the LLMs, conservative for `bibtex-updater`), with coverage beside it; Tab. 25 gives all three. The conservative–aggressive gap is the cost of the abstention, and it is non-trivial exactly where coverage is below one: `bibtex-updater` on `dev_public` (coverage 0.86) earns selective DR/FPR/F1 of 0.992/0.056/0.971 on the 965 entries it commits to, while over the full split conservative scoring (abstain→VALID) gives 0.820/0.051/0.880 and aggressive scoring (abstain→HALLUCINATED) gives 0.987/0.140/0.937; DeepSeek-R1 on `test_public` (coverage 0.783) pays +16.2 pp between its conservative and aggressive stances. The selective FPR (0.056) essentially matches the full-coverage headline FPR (0.051): abstention does not add precision for `bibtex-updater`; it raises recall (DR 0.820 →0.992) at near-unchanged FPR. Reporting Coverage *with* the aggressive number closes the obvious loophole—abstaining on everything would otherwise give perfect committed-precision—so a tool cannot hide a weak verdict behind UNCERTAIN.

Risk–coverage. For each tool with graded confidences we abstain on the least-confident entries first (smallest $|P(\text{hallucinated}) - 0.5|$) and score FPR on the rest, tracing a risk–coverage curve (data in the released `coverage_reporting.json`; we summarize each curve by its FPR at 90% coverage in Tab. 25b). Abstaining on the least-confident decile yields a real FPR reduction for the better-calibrated LLM verifiers: the high-AUROC verifiers (Sonnet 4.6, Opus 4.7; AUROC 0.90–0.93) sit at the favorable corner with FPR ≈0.10–0.12 at 90% coverage, while the recall-aggressive open-weight models stay above 0.5. `bibtex-updater` is the opposite case: its risk–coverage

Table 25: **Coverage and selective prediction on `dev_public`.** *(a)* Per-tool coverage with selective (abstentions excluded) and aggressive (abstain→HALLUCINATED) DR/FPR/F1; conservative scoring (abstain→VALID) is the convention of Tab. 1. On `test_public`, `bibtex-updater`'s conservative and aggressive stances are 0.838/0.080/0.889 and 0.990/0.151/0.952, and its selective DR/FPR/F1 is 0.989/0.086/0.967 (n=730). *(b)* FPR at 90% coverage from the risk–coverage curve; − marks tools the threshold sweep did not cover (curves are in the released JSON). In (b), blue marks FPR $\leq$.12 (the favorable risk–coverage corner) and red FPR $>$.5; the gray row in (a) is the co-designed reference.

(a) Coverage + dual scoring

Tool	Cov.	Selective DR/FPR/F1	Aggr. DR/FPR/F1
`bibtex-updater` (v1.2.0)	0.86	0.992 / 0.056 / 0.971	0.987 / 0.140 / 0.937
Gemini 2.5 Pro	0.967	0.476 / 0.050 / 0.627	0.497 / 0.074 / 0.637
Gemini 2.5 Flash	0.988	0.500 / 0.100 / 0.631	0.508 / 0.105 / 0.636
Mistral Large	0.989	0.716 / 0.251 / 0.742	0.721 / 0.253 / 0.745
DeepSeek-R1	0.984	0.896 / 0.623 / 0.739	0.898 / 0.628 / 0.739
Opus 4.7	1.000	0.906 / 0.154 / 0.890	0.906 / 0.154 / 0.890
Sonnet 4.6	0.982	0.916 / 0.161 / 0.893	0.918 / 0.177 / 0.888

(b) FPR at 90% coverage (risk–coverage summary)

Tool	AUROC	FPR@90%	F1@90%
Sonnet 4.6	0.927	0.119	0.912
Opus 4.7	0.903	0.098	0.903
GPT-5.4	0.834	0.222	0.816
Mistral Large	0.744	0.241	0.736
Gemini 2.5 Flash	0.739	0.097	0.578
DeepSeek-R1	0.741	0.581	0.752
DeepSeek-V3.2	0.609	0.699	0.711

curve is flat—FPR is essentially constant (0.092 at full coverage, 0.099 on the committed subset)—so selective prediction adds no precision. Where the LLM verifiers lower their false-positive rate as they abstain, `bibtex-updater`'s precision is structural; it is the precision-anchored reference against which the LLMs' abstention gains are read.

Two caveats. (i) LLM abstention is prompt-dependent (§ H.2: Sonnet's UNCERTAIN rate moves 2.0% → 8.7% by wording), so coverage is read as "coverage under the default prompt", not an intrinsic constant. (ii) `bibtex-updater`'s abstention and an LLM's occupy the same column but differ in mechanism: the tool abstains because *the databases hold no matching record* (a data-coverage gap), whereas an LLM abstains because *it will not commit* (an epistemic gap); under the route-to-human framing both are the same useful action, which is why the column is shared. `bibtex-updater` abstains on 199/1119 entries on `dev_public` (coverage 0.822) and 171/831 on `test_public` (coverage 0.794).

Takeaway. The precision-ceiling finding survives the obvious confounds: rankings are invariant to prompt wording (§ H.2) and decision threshold (§ H.4), detection leans on title/author content rather than a format tell (§ H.5), and independent raters reproduce the over-flagging the benchmark measures (§ H.6). Coverage makes the abstention lever explicit; the better-calibrated LLMs trade a small coverage cut for a real FPR reduction (precision-via-abstention), whereas the co-designed `bibtex-updater`'s low FPR (0.092) is structural: its risk–coverage curve is flat, so abstaining on the 12–13% of entries it cannot verify recovers recall (DR 0.865 → 0.987) at near-unchanged FPR rather than adding precision.

E.3 Regime-conditional deployment guidance

When to prefer recall. The operating point is set by three quantities: prevalence, the cost ratio $c_{\text{FN}}/c_{\text{FP}}$ of a missed fabrication versus a false alarm, and reviewer capacity. FPR-ranking holds in the *reviewer-bound venue-audit* regime—low prevalence, finite reviewer attention, a false alarm cheap to dismiss—where precision is the limiting factor. When a missed fabrication is far costlier than triaging extra flags, or when a later human-review stage filters out the false alarms the tool raises, recall plus human triage dominates and a high-recall tool (an agentic harness, or DeepSeek-V3.2) is preferable despite its FPR. The two takeaways—"rank on FPR" and "a missed hallucination is the worse error"—are not in tension; they name different aspects of the (prevalence, $c_{\text{FN}}/c_{\text{FP}}$, capacity) space, which Tab. 26 resolves into tool choices.

Table 26: **Regime-conditional deployment guidance.** The right verifier depends on prevalence, the cost of a missed fabrication relative to a false alarm ($c_{\text{FN}}/c_{\text{FP}}$), and reviewer capacity; per-tool numbers in § E.3.

Regime	Pick	Why
Reviewer-bound venue audit—low prevalence, a false alarm cheap to dismiss	Low-FPR verifier with usable recall (`bibtex-updater`); among LLMs, the best-calibrated Sonnet 4.6 or Opus 4.7	At low prevalence false alarms outnumber true detections even for the best verifier (precision $\sim$16% at a 2% base rate), so a few points of FPR decide usability; add the Stage-2 cascade when recall must rise at the same FPR.
High prevalence—triage queues, pre-submission self-check	High-recall verifier (an agentic harness or DeepSeek-V3.2)	Precision improves as the base rate rises, so catching more matters and the extra false alarms are affordable.
Costly miss, or downstream human review absorbs false alarms	High-recall verifier + human triage	When a fabrication in print far outweighs an extra flag, recall plus a human filter is a better choice.

At venue-realistic prevalence ($\sim$2%), the precision-bottleneck logic favors low-FPR tools regardless of recall: PPV per flagged entry is $\sim$16% for Gemini 2.5 Pro and the co-designed `bibtex-updater`, $\sim$10–11% for Opus 4.7 and Sonnet 4.6, $\sim$4% for GPT-5.1, and $<$3% for high-recall open-weight models (Tab. 23).

- **Pre-2024, recall-prioritized triage:** a high-recall verifier such as an agentic harness (DR $\approx$0.97–0.99) or DeepSeek-V3.2 (DR = 0.911), accepting the higher FPR where reviewer capacity absorbs false positives. **Pre-2024, precision-prioritized:** `bibtex-updater` (DR = 0.820, FPR = 0.051, PPV $\sim$25% at a 2% base rate, $\sim$2–3 orders of magnitude cheaper than LLM verifiers), or, among LLMs, Sonnet 4.6 or Opus 4.7 (FPR = 0.161 / 0.154, DR above 0.9); Gemini 2.5 Pro has lower FPR (0.050) but misses half the hallucinations.
- **Near or beyond training cutoffs:** Sonnet 4.6 zero-shot for the best calibrated FPR among LLMs tested. Agentic harnesses match higher detection but inflate FPR over their zero-shot bases (GPT-5.1 +6 pp; Sonnet 4.6 +27 pp, a 2.7$\times$ multiplier from a low base); deploy only where reviewer capacity absorbs false positives.
- **Past 12 months:** no tool we tested is reliable without cutoff-aware prompting (§ F.2); the addendum recovers FPR on GPT-5.1 and Qwen3-235B at the cost of indiscriminate abstention, barely moves Gemini 2.5 Flash, and abstains selectively only on Sonnet 4.6, so it requires per-provider tuning.

F Failure mode (iii): temporal fragility and calibration

Following LiveCodeBench (Jain et al., 2025), we tag entries with three temporal segments (*pre-2023*, *2023–2024*, *2025+*) so the benchmark can measure how verifiers degrade on recent papers (§ 4.4). We read that degradation through the cross-regime *ranking*, which is prompt-invariant, rather than through absolute post-cutoff FPR, which a prompt-sensitivity sweep moves by 10–37 percentage points (pp) on wording alone (§ H.2).

Training-data cutoffs of evaluated LLMs. The temporal evaluation assumes papers from calendar years 2024–2025 post-date most evaluated models' training data. Tab. 27 compiles vendor-disclosed training cutoffs (or the most recent reliable upper bound when no disclosure is available) for every LLM used as a baseline. Three observations frame the temporal results. (i) Every evaluated model has a disclosed or inferred cutoff no later than late 2025, so the 448-entry temporal supplement (drawn from 2024–2025 DBLP proceedings) is partially or fully out-of-distribution for all of them. (ii) The two closed-weight models with earliest cutoffs (GPT-5.1, Sep 2024) and the conservative open-weight Gemini 2.5 Flash (Jan 2025) are hardest-hit by the post-cutoff FPR rise, consistent with their reliance on parametric confirmation. (iii) DeepSeek models do not publish an official cutoff in their technical reports, so we report the OpenRouter-surfaced metadata and flag it as a provenance gap that future versions of HALLMARK should track via release-date probes.

2024–2025 DBLP papers are out-of-distribution to varying degrees across the cohort: late-2025 entries are post-cutoff for every model, while early-2024 entries are post-cutoff only for the earliest-cutoff models (GPT-5.1: Sep 2024; DeepSeek-R1: Jul 2024). Across the nine non-Anthropic LLMs tested on the supplement the temporal hit splits into four regimes: four models with healthy baselines degrade sharply (GPT-5.1, Gemini 2.5 Flash, Mistral Large, Llama 4 Maverick; ΔFPR +35–62 pp);

Table 27: **Training-data cutoffs of LLM baselines evaluated in HALLMARK.** "Disclosed" indicates a vendor-published cutoff in an official model card, API documentation, system card, or technical report; "Inferred" indicates an upper bound derived from release date and predecessor disclosures. Anthropic distinguishes "training data cutoff" from "reliable knowledge cutoff"; we report the latter for Claude variants. Rows above the midrule have cutoffs at or before mid-2025; rows below have post-Aug 2025 cutoffs (the "later-cutoff" cohort referenced in § 4.4).

Model	Training cutoff	Source	Status
DeepSeek-R1	Jul 2024	OpenRouter model metadata	Disclosed (via host)
Llama 4 Maverick	$\leq$ Aug 2024	Meta release notes; release Apr 2025	Inferred
GPT-5.1	Sep 2024 (GPT-5 family)	OpenAI system card (GPT-5)	Disclosed (family)
DeepSeek-V3.2	$\leq$ Oct 2024	Release date; V3 tech report	Inferred
Gemini 2.5 Flash	Jan 2025	Vertex AI model documentation	Disclosed
Gemini 2.5 Pro	Jan 2025	Vertex AI model documentation	Disclosed
Qwen3-235B-A22B-2507	Jun 2025	OpenRouter model metadata	Disclosed (via host)
Qwen3-VL-235B-A22B-Instruct	$\leq$ Jun 2025	OpenRouter metadata; Qwen3 family lag	Inferred
Mistral Large 2512	$\leq$ mid-2025	Release date; predecessor cutoff	Inferred
GPT-5.4	Aug 2025	OpenAI release (`gpt-5.4-2026-03-05`)	Disclosed
Claude Sonnet 4.6	$\leq$ Aug 2025	Release Feb 2026; Anthropic family lag	Inferred
Claude Opus 4.7	$\leq$ Oct 2025	Release Apr 2026; Anthropic family lag	Inferred

three with already-elevated baselines stay near-ceiling or worsen (Qwen3-235B, DeepSeek-V3.2, Qwen3-VL-235B at FPR 0.89, a new cohort maximum); DeepSeek-R1's run failed on 151 of the 448 entries, leaving only its FPR of 0.856 evaluable (Tab. 28, footnote †); and Gemini 2.5 Pro over-flags only moderately (FPR 0.25). Pro-tier post-training, not data recency, drives that last case: Gemini 2.5 Flash—same vendor, comparable cutoff—degrades sharply (FPR 0.595). Training-data recency alone does not rescue parametric verification. The latest-cutoff non-Anthropic model tested (Qwen3-235B: Jun 2025) still ends in the near-ceiling group, so the gap between "seen during training" and "reliably recalled" matters as much as the cutoff. The two later-cutoff Anthropic models (Sonnet 4.6, Opus 4.7) sharpen the point: their inferred cutoffs (Aug 2025, Oct 2025) are not radically later than Qwen3-235B's, yet their FPR shifts on the supplement are an order of magnitude smaller. Cutoff recency is therefore necessary but insufficient; pipeline-level differences (Reinforcement Learning from Human Feedback (RLHF) for abstention, "I don't know" calibration) appear to dominate the temporal-robustness story.

Baseline temporal consistency. For bibtex-updater, detection rates are consistent across temporal segments ($\pm 2\%$); DOI-only improves slightly on newer entries because recent papers more consistently include DOIs.

F.1 TEMPORAL SUPPLEMENT (2024–2025)

The pattern extends past 2024–2025: a 60-entry probe on 2026 arXiv submissions reproduces the FPR multiplier ($r = -0.82$ between baseline aggressiveness and relative post-cutoff degradation).
To validate the 60-entry temporal probe at scale (probe composition in § A.3), we construct a 448-entry temporal supplement: 300 valid entries scraped from DBLP across six ML venues (NeurIPS, ICML, ICLR, AAAI, CVPR, ECCV) for publication years 2024–2025, plus 148 hallucinated entries generated via the standard perturbation pipeline (58/52/38 across Tiers 1/2/3, spanning all 14 types; Tab. 5). This provides $10\times$ more valid entries than the probe (300 vs. 30), yielding substantially tighter confidence intervals on FPR estimates. The 448-entry set (300 valid / 148 hallucinated, full coverage) is the canonical temporal supplement reported here; every released prediction file covers the same 448 keys, several of them also carrying rows from the 858-entry superset. A later 858-entry superset of the same 2024–2025 DBLP pool backs the extended reviewer-experiment robustness checks (recall probe, late-cutoff control, GPT-5.1 run-to-run variance) reported separately in the released artifacts (`results/reviewer_experiments/FINDINGS.md`), and is presented alongside—not in place of—the canonical 448-entry table.
Tab. 28 compares each LLM baseline's performance on the original benchmark (2021–2023 valid entries) against the temporal supplement (2024–2025 valid entries). The pattern is unambiguous: every model's FPR increases dramatically on recent papers, while detection rate inflates because models flag nearly everything as hallucinated.

†DeepSeek-R1's run on this supplement returned 151 of 448 entries as `[Error fallback]` API failures recorded as `UNCERTAIN`, and those 151 include all 148 hallucinated entries. Its detection rate, ΔDR, F1 and MCC therefore measure the harness rather than the model and we mark them not evaluable (n/e). The FPR of 0.856 is scored over the 297 valid entries the run did answer and is comparable to the other rows.

Table 28: **Temporal supplement: LLM baseline performance on 2021–2023 benchmark entries (`dev_public`, n=1,119) vs. 2024–2025 supplement entries (n=448).** ΔDR and ΔFPR report the absolute shift. The top block (Gemini Flash through DeepSeek-R1) covers the six LLMs with training cutoffs at or before mid-2025, sorted by baseline FPR ascending; DeepSeek-R1 is placed last because its 24–25 behavior is anomalous (see †). Below the midrule are five models added later (2026-05-05 collection; § B.2.4): the two later-cutoff Anthropic models (Sonnet 4.6, Opus 4.7; documented or inferable post-Aug 2025 cutoffs, § F) do not show the post-cutoff FPR rise, while Gemini 2.5 Pro over-flags only moderately and Llama 4 Maverick and Qwen3-VL-235B degrade like the top block. †DeepSeek-R1's detection rate on this supplement is not evaluable; see footnote below table. In the $FPR_{24–25}$ column, red marks the over-flagging cluster (FPR $\geq$.59) and blue the two models whose post-cutoff FPR holds; Gemini 2.5 Pro (.250) sits between and is unshaded. **Snapshot.** Every row pairs both columns from one run: a Δ across the cutoff is only interpretable within a single endpoint snapshot. The Sonnet 4.6 and Opus 4.7 rows therefore stay on the May snapshot, so their 21–23 figures (.127, .072) are the May ones and differ from the later 2026-09-02 collection reported in Tab. 1 (.161, .154); Tab. 29 quantifies that gap.

Model	$DR_{21–23}$	$DR_{24–25}$	ΔDR	$FPR_{21–23}$	$FPR_{24–25}$	ΔFPR	$F1_{24–25}$	$MCC_{24–25}$
Gemini Flash	0.500	0.844	+0.344	0.100	0.595	**+0.495**	0.555	0.251
GPT-5.1	0.837	0.958	+0.121	0.411	0.759	**+0.348**	0.546	0.246
Mistral Large	0.716	0.952	+0.236	0.250	0.793	**+0.543**	0.537	0.208
Qwen3-235B	0.860	0.986	+0.126	0.533	0.809	+0.276	0.541	0.245
DeepSeek-V3.2	0.911	0.980	+0.069	0.702	0.759	+0.057	0.558	0.278
DeepSeek-R1†	0.896	n/e	n/e	0.623	0.856	+0.233	n/e	n/e
Later-added models (2026-05-05 collection):								
Claude Sonnet 4.6	0.781	0.818	+0.037	0.127	**0.120**	−0.007	0.793	0.688
Claude Opus 4.7	0.752	0.716	−0.036	0.072	**0.073**	+0.001	0.768	0.669
Gemini 2.5 Pro	0.476	0.626	+0.150	0.050	0.250	+0.200	0.588	0.366
Llama 4 Maverick	0.614	0.939	+0.325	0.146	0.763	**+0.617**	0.539	0.216
Qwen3-VL-235B	0.860	1.000	+0.140	0.551	0.887	+0.336	0.527	0.201

Table 29: **Endpoint drift for the two Anthropic models on `dev_public`.** The same model ID, prompt, split, temperature and seed, collected on two dates: **May** = 2026-05-05, **Sep** = 2026-09-02 (§ B.2.4). Tab. 28 uses the May column so that each of its rows is within-snapshot; Tab. 1 reports the September column as the later run. Δ is September minus May. A controlled replay on shared inputs gives only 90%/75% Opus/Sonnet per-entry label agreement between the two collections, so these are distinct measurement sessions rather than repeated draws from one. The F1 columns show why we decline to rank these two models against each other: the gap between them is 0.3 pp in both collections, but its *sign* reverses (bold marks the leader), so which model leads is a property of the session rather than of the models.

	DR			FPR			F1	
Model	May	Sep	Δ	May	Sep	Δ	May	Sep
Claude Sonnet 4.6	0.781	0.916	+0.135	0.127	0.161	+0.034	0.827	**0.893**
Claude Opus 4.7	0.752	0.906	+0.154	0.072	0.154	+0.082	**0.830**	0.890

Endpoint drift between the two collections. The Sonnet 4.6 and Opus 4.7 rows above are read off the 2026-05-05 collection, while Tab. 1 reports the 2026-09-02 collection. Tab. 29 gives both. The two runs share prompt, split, temperature and seed, so the difference is attributable to the endpoint behind a fixed model name rather than to any change we made. The size of that drift is why the rows above are not mixed: on `dev_public` the FPR moves +0.034 (Sonnet) and +0.082 (Opus) between collections, which is larger than the post-cutoff shift the table is built to measure (−0.007 and +0.001). Pairing a September 21–23 figure with a May 24–25 figure would therefore report drift as if it were a temporal effect, and with the signs as they fall here it would read as FPR *improving* on post-cutoff papers (−0.041, −0.081).

The supplement confirms three patterns from the probe, now with substantially greater statistical power, and surfaces a fourth via the later-evaluated models. First, **FPR rises sharply on recent papers for most models**: GPT-5.1's FPR climbs from 41.1% to 75.9% (1.8×); three conservative-baseline models (Gemini Flash +49.5 pp, Mistral Large +54.3 pp, Llama 4 Maverick +61.7 pp)

post the largest absolute jumps, while GPT-5.1's increase is more modest at +34.8 pp. Second, **the detection-rate increases are spurious**: the four largest-jump models (Gemini Flash, Mistral Large, Llama 4 Maverick, GPT-5.1) achieve DR $\geq$0.84 on the supplement, but only because they flag nearly everything; F1 falls to $\approx$0.55 (from 0.61–0.77 on `dev_public`) and MCC to $\approx$0.25, confirming near-random predictions. Third, **already-aggressive models stay near-ceiling**: DeepSeek-V3.2 (baseline FPR 70.2%) shows only a 5.7 pp increase but ends at 75.9%; Qwen3-VL-235B is the most extreme, climbing 55.1% $\rightarrow$ 88.7%, a new cohort maximum. Fourth, **a vendor-/pipeline-correlated subset resists**: Claude Sonnet 4.6 (FPR 12.0%) and Opus 4.7 (7.3%) hold; Gemini 2.5 Pro (FPR 25.0%) and a later-cutoff GPT-5.4 (FPR 41.3%) over-flag only moderately: pipeline differences appear to dominate over training recency for the residual gap.

The 60-entry probe, which additionally covered 2026 arXiv submissions, showed the same FPR-multiplier pattern with a strong negative correlation ($r = -0.82$) between baseline aggressiveness and relative degradation, evidence that the failure mode generalizes beyond 2024–2025 DBLP.

These results establish that LLM citation verification leans heavily on parametric knowledge with a graded temporal boundary near the training cutoff: most models fall into the "flag everything unfamiliar" failure mode for post-cutoff papers. But the partial-to-full resistance from Gemini 2.5 Pro, GPT-5.4, and the two Anthropic models shows this is not a hard structural limit of parametric verification: training-pipeline interventions (post-training calibration, abstention RLHF) can shift the trade-off without retrieval augmentation, even at comparable training cutoffs. We caution against reading the latest-cutoff models' low post-cutoff FPR as evidence of better *calibration*: the 2024–2025 supplement is drawn from DBLP proceedings, so for a model whose training data plausibly ingested those same DBLP records, a correct VALID verdict is indistinguishable from training-data *recall* (contamination) rather than principled abstention or verification. We therefore do not attribute their resistance to calibration on this supplement alone: the recall probe and the third-provider control (§ F.4) point toward calibration for Opus 4.7 without excluding recall, and separating the two fully would require post-cutoff valid entries the models provably never saw, which the current supplement cannot guarantee.

Takeaway. The post-cutoff failure is graded, not binary: 8 of the 12 evaluated LLMs over-flag on 2024–2025 papers (FPR $\approx$0.60–0.89 for the seven near-ceiling models plus Gemini Flash at 0.595), while Sonnet 4.6 (0.120) and Opus 4.7 (0.073) hold and Gemini 2.5 Pro over-flags only moderately (0.250). The 12th LLM, the later-cutoff GPT-5.4, is evaluated in the separate temporal-cutoff probe (§ F.3) and over-flags only moderately (supplement FPR 0.41). Training-pipeline differences (abstention RLHF, calibration) outweigh recency; don't trust a model just because its cutoff is recent. Caveat: n=2 Anthropic models.

F.2 Robustness check: cutoff-aware prompting

S2, Crossref and OpenAlex exhibit weeks-to-months ingestion lag for new arXiv preprints, contributing additively to the post-cutoff FPR rise, while the arXiv API itself covers preprints immediately but not venue-published versions; the empirical confirmation follows.

Mitigation and controls. A cutoff-aware prompting variant (§ F.2) recovers most of GPT-5.1's post-cutoff FPR (75.9%$\rightarrow$0.0% on the entries it still commits to) but inflates its abstention uniformly (pre-cutoff UNCERTAIN rises to 52.7%), whereas on Sonnet 4.6 the same addendum abstains selectively (post-cutoff UNCERTAIN 48.9% against 8.7% pre-cutoff) while halving its committed-entry FPR; the failure is epistemic miscalibration, not structural blindness, and the addendum's effect is model-dependent. The later-cutoff GPT-5.4 control shows that recency helps but does not cure over-flagging: on these 2024–2025 papers, inside its Aug 2025 window, its FPR is 41.3% (§ F.3), well below the earlier-cutoff over-flagging cluster yet far above the latest-cutoff models.

The temporal results above establish **H1**: LLMs do not know their own cutoff; post-cutoff FPR rises sharply because the model treats unfamiliar papers as suspicious. A separate hypothesis remains untested: **H2**, that LLMs can route post-cutoff citations to `UNCERTAIN` *if explicitly reminded of the cutoff*. H1 and H2 are orthogonal: H1 failing does not imply H2 must fail (the model may have latent metacognitive capacity that the default prompt does not elicit), and H2 failing would upgrade H1 from "epistemic miscalibration" to "structural blindness to temporal uncertainty."

Prompt addendum. Each cutoff-aware variant appends the following text to the default verification prompt, verbatim:

```
Note: your training data has a knowledge cutoff. If the citation could
post-date your training data, or if you cannot recall the paper with
confidence, respond with UNCERTAIN rather than HALLUCINATED or VALID.
Do not guess.
```

No other element of the prompt, temperature (T=0), seed (42), or token budget (`max_completion_tokens=1024`, the non-Anthropic value of § B.2.4; Sonnet 4.6 runs through the OpenRouter mirror, which uses the same budget) is changed, so any observed shift in predictions is attributable to the addendum alone.

Subset of models. We run the cutoff-aware variant on GPT-5.1, Gemini 2.5 Flash, Qwen3-235B, and Claude Sonnet 4.6. DeepSeek-R1 is omitted because its default-prompt run on this supplement failed on 151 of 448 entries (Tab. 28, footnote †), leaving no usable default-prompt baseline to compare an addendum against; DeepSeek-V3.2 is omitted because it already sits at the aggressive ceiling (FPR 0.759 with DR 0.980), where any measured "improvement" would be dominated by noise. Mistral Large is omitted because its FPR dynamics closely track Qwen3-235B. The Sonnet 4.6 sweep was added later (2026-07-02, through the OpenRouter mirror) with both a cutoff-aware and a same-snapshot default pass on the canonical 448-entry supplement and the 150-entry pre-cutoff sample, so its comparison is fully within-run and unaffected by the endpoint drift of § 5. The four-model subset spans the full conservative–aggressive FPR spectrum observed in Tab. 28: Sonnet 4.6 anchors the calibrated end, Gemini Flash and GPT-5.1 populate the conservative-to-middle range, and Qwen3-235B the aggressive end.

Metrics. We report FPR and the `UNCERTAIN` rate (coverage is its complement) per segment, stratified by each model's training-data cutoff (Tab. 27) into pre-cutoff and post-cutoff entries. The comparison is computed from cached predictions via `hallmark.evaluation.temporal.compare_prompt_variants`, so the ablation is reproducible from the released prediction dumps without re-running the API calls. The default-prompt FPR baseline here (FPR_{def} in Tab. 30) matches the main temporal supplement (Tab. 28) for GPT-5.1, Gemini 2.5 Flash and Qwen3-235B, scored on the same 448 entries; Sonnet 4.6 uses its own 2026-07-02 snapshot (18.6% vs. 12.0% there), so read its ΔFPR within this experiment.

Table 30: **Cutoff-aware ablation results.** *Post-cutoff*: 2024–2025 temporal supplement (N=448 unique entries, all post-cutoff for the four tested models). *Pre-cutoff*: 150-entry stratified sample of DEV_PUBLIC (2021–2023, pre-cutoff for all four models; 75 VALID/75 HALLUCINATED). Default pre-cutoff `UNCERTAIN` rates are taken from the full DEV_PUBLIC evaluation for the top three models; the Sonnet 4.6 rows (below the midrule) were collected on a later dated snapshot (2026-07-02) with a same-snapshot default baseline on both pools, so its comparison is fully within-run (its default supplement FPR reads 18.6 here against 12.0 in Tab. 28, the span of the endpoint drift). Δ columns report cutoff-aware minus default. Red marks the mitigation's cost: pre-cutoff abstention inflated by the addendum, so the FPR recovery is not a free deployment win.

	Post-cutoff (N=448)				**Pre-cutoff** (N=150)	
Model	FPR_{def}	FPR_{CA}	ΔFPR	UNC_{CA}	UNC_{def}	UNC_{CA}
GPT-5.1	75.9	0.0	-75.9	94.4	0.0	52.7
Gemini 2.5 Flash	59.5	47.0	-12.5	19.2	1.2	0.7
Qwen3-235B	80.9	0.0	-80.9	99.8	0.2	95.3
Claude Sonnet 4.6	18.6	9.7	-8.9	48.9	3.3	8.7

Outcome classification (per model). The four tested models exhibit qualitatively different outcomes, so a single aggregate classification would obscure the signal.

- **GPT-5.1**: FPR drop = 75.9 pp, pre-cutoff `UNCERTAIN`= 52.7%, post-cutoff `UNCERTAIN`= 94.4% → *mostly blanket abstention*: it abstains on 94% of new and 53% of old papers.
- **Gemini 2.5 Flash**: FPR drop = 12.5 pp, pre-cutoff `UNCERTAIN`= 0.7%, post-cutoff `UNCERTAIN`= 19.2% → *strong H1 confirmation*.
- **Qwen3-235B**: FPR drop = 80.9 pp, pre-cutoff `UNCERTAIN`= 95.3%, post-cutoff `UNCERTAIN`= 99.8% → *blanket abstention*: `UNCERTAIN` on 95–100% of entries, old or new, so the FPR drop is not selective.
- **Claude Sonnet 4.6**: FPR drop = 8.9 pp (18.6% → 9.7% on committed entries), pre-cutoff `UNCERTAIN`= 8.7%, post-cutoff `UNCERTAIN`= 48.9% → *selective abstention, the closest to H2 in the cohort*.

Sonnet 4.6 comes closest to *H2 holds*: it is the only model whose abstention inflates selectively—post-cutoff `UNCERTAIN` rises to 48.9% while pre-cutoff abstention stays below 9%—and its committed-entry FPR halves (18.6% → 9.7%). The selectivity still routes roughly half the post-cutoff entries

to a human, so abstention remains the mitigation's price. The other three respond in qualitatively different ways: GPT-5.1 and Qwen3-235B lose discrimination (post-cutoff UNCERTAIN above 94%, with pre-cutoff abstention inflating by 52 pp and 95 pp respectively), while Gemini 2.5 Flash is barely moved by the reminder (FPR drop only 10.5 pp while pre-cutoff UNCERTAIN stays near zero). The practical implication stands: prompt-level temporal-awareness mitigations are *not portable across models*—the four responses differ qualitatively—and cannot be relied on as a substitute for retrieval-augmented verification.

Takeaway. The cutoff-aware addendum confirms epistemic miscalibration over structural blindness: on GPT-5.1 it drops post-cutoff FPR from 75.9% to 0.0% but inflates pre-cutoff UNCERTAIN to 52.7%, trading discrimination for abstention; Sonnet 4.6 is the only model that abstains selectively (post-cutoff UNCERTAIN 48.9% against 8.7% pre-cutoff) while halving its committed-entry FPR (18.6% → 9.7%). Prompt-level mitigation works yet isn't a free deployment win. Caveat: n=4 models tested.

F.3 Later-cutoff robustness check: GPT-5.4 on the temporal supplement

H1 predicts that the FPR rise on post-cutoff entries is a *direct* consequence of the training-data cutoff: moving the cutoff forward should shift which entries the model recognizes and, in turn, drop the FPR on entries that cross the new cutoff line. We probe this with a single later-cutoff model: **GPT-5.4** (`gpt-5.4-2026-03-05`), released on 5 March 2026 with an **August 31, 2025** training cutoff, roughly 11 months later than GPT-5.1's September 2024 cutoff. We run it zero-shot on the same 448-entry 2024–2025 temporal supplement used in Tab. 30 and compare entry-by-entry against the existing GPT-5.1 default-prompt predictions. GPT-5.4 is reported here as a single-model temporal-cutoff probe; its zero-shot `dev_public` performance is also included as a baseline row in Tab. 1. The GPT-5.4 per-entry predictions on all 448 supplement entries are released, so the probe's aggregate FPR (0.413) and DR (0.899) are reproducible offline without further API calls.

Table 31: **Stratified GPT-5.1 vs. GPT-5.4 on the 2024–2025 temporal supplement (N=448).** Stratum is the entry's publication year. *2024 & earlier* (236 entries: 161 valid, 75 hallucinated) is post-cutoff for GPT-5.1 but pre-cutoff for GPT-5.4. *2025* (191 entries: 139 valid, 52 hallucinated) is post-cutoff for both, although some venues (ICLR/ICML/AAAI/CVPR/ACL 2025) predate August 2025. *Future* (21 entries: all hallucinated `future_date` types) is post-cutoff for both and contains no valid entries, so FPR is undefined. GPT-5.1 cells are computed from its released per-entry predictions matched to the canonical 448 entries with abstentions excluded (9/448), consistent with Tab. 28. The red-shaded cell is the residual: even on pre-cutoff entries GPT-5.4 over-flags 28% of valid papers, so a later cutoff alone does not close the gap.

		FPR (%)			DR (%)	
Stratum	N	GPT-5.1	GPT-5.4	Δ	GPT-5.1	GPT-5.4
2024 & earlier	236	56.2	28.0	−28.2	92.0	85.3
2025	191	99.3	56.8	−42.4	100.0	92.3
Future (2026+)	21	–	–	–	100.0	100.0
All	448	75.9	41.3	−34.6	95.9	89.9

Three features of the result replicate the paper's temporal narrative across a model generation:

- **FPR drops monotonically with cutoff distance.** The 2024-and-earlier stratum shifts from post-cutoff (GPT-5.1) to pre-cutoff (GPT-5.4) and FPR drops by 28 pp. The 2025 stratum is post-cutoff for *both* models but is *closer* to the GPT-5.4 cutoff; FPR drops by 42 pp, consistent with a graded notion of "unfamiliar" rather than a binary cutoff.
- **GPT-5.4 still over-flags.** On pre-cutoff 2024 entries, GPT-5.4's FPR is 28%, far from zero. Training cutoff alone does not explain the full FPR burden: long-tail recognition, paraphrased titles, and pre-print/published-version mismatches contribute residual false positives even for papers the model's training data should have covered. This tempers a naïve "just use a newer model" mitigation.
- **UNCERTAIN rate does not rise.** Unlike the cutoff-aware prompt ablation (Tab. 30), which inflated abstention without shifting discrimination, moving to a later-cutoff model reduces FPR *and* drops UNCERTAIN from 2.0% to 0.0%. The gain comes from familiarity with the newer entries rather than from added caution.

We interpret this as evidence that H1 is a genuine temporal phenomenon rather than an artifact of GPT-5.1 specifically, and that the practical mitigation path is retrieval-augmented verification rather than prompt-engineered abstention, consistent with the agentic-baseline finding in Tab. 1.

Takeaway. H1 replicates across model generations: GPT-5.4's later cutoff drops FPR by 28–42 pp stratum-wise (75.9%→41.3% overall) and UNCERTAIN falls to 0%: familiarity, not caution. Yet pre-cutoff FPR is still 28%, so "deploy a newer model" does not close the gap; pair it with retrieval-augmented verification.

F.4 CALIBRATION-VS-MEMORIZATION CONTROLS

Two controls push against, but do not eliminate, the memorization reading (§ F.4). A recall probe finds Opus 4.7 accepts 84% of valid 2024–2025 papers it cannot recall, consistent with calibration rather than memorization; the recall measure is model-derived—we count only papers the model itself reports it cannot recall—so contamination is reduced but not excluded. Sonnet 4.6's lower FPR is partly recall-driven. A third-provider late-cutoff control (DeepSeek-V4-Pro, complete n=300 run at full coverage) holds its post-cutoff FPR at 0.36 where GPT-5.1 reaches 0.93 on the same entries, consistent with the effect spanning providers without establishing it.[4] The resistance points away from pure memorization for Opus 4.7, but with N=2 holding models from one provider and a single third-party control, it remains a preliminary signal that stops short of a causal attribution.

Three targeted experiments probe the mechanism behind the temporal-robustness results, using the temporal supplement and a `dev_public` subsample. Each is reported with its own N and complements Tab. 28. All three were run on 2026-05-28 through the HALLMARK Python API: GPT-5.1 via the native OpenAI endpoint (`gpt-5.1`), and the Anthropic and DeepSeek models via OpenRouter (`anthropic/claude-sonnet-4.6`, `anthropic/claude-opus-4.7`, `deepseek/deepseek-v4-pro`).

(E1) Recall probe: calibration vs. memorization. A low post-cutoff FPR on a real paper is ambiguous: the model may be *calibrated* (declining to flag the unfamiliar) or merely *recalling* a paper its training data contained. This is the contamination confound noted in § 4.4: the supplement's valid entries are scraped from DBLP, and a model whose training data ingested those same DBLP records can accept them from memory rather than verification. On 150 valid 2024–2025 papers we run, per model, the standard verifier (FPR) and a recall probe that supplies only the title and year and asks for the author list and venue with no lookup ("recalled" = predicted-author last-name Jaccard ≥ 0.5 vs. truth). If a low FPR were memorization-driven, accepted-as-valid verdicts would concentrate on recalled papers.

Table 32: **Recall probe on 150 valid 2024–2025 papers.** $P(\text{VALID} \mid \text{recalled})$ vs. $P(\text{VALID} \mid \text{not recalled})$ separates memorization from calibration. Verify FPR is flagged/150; UNC entries are neither flagged nor accepted.

Model	Recall	Verify FPR	UNC	$P(\text{v} \mid \text{rec})$	$P(\text{v} \mid \overline{\text{rec}})$
GPT-5.1	0%	92%	0%	—	8%
Claude Sonnet 4.6	10%	24%	23%	87%	50%
Claude Opus 4.7	29%	13%	1%	91%	**84%**

Opus 4.7 accepts 84% of valid papers it *cannot* recall, so its low FPR reflects genuine calibration rather than memorization; Sonnet 4.6 accepts 87% of papers it recalls against 50% of those it cannot, but on the latter it abstains on 24% rather than flagging them, so part of the gap is abstention, not memorization; GPT-5.1 recalls none and flags 92%. This *partially* resolves the contamination confound: it is largely ruled out for Opus 4.7 and only partly for Sonnet 4.6. Caveats: title-given recall is a lower bound on memorization; abstention and "cannot recall" are not perfectly separable; N=150, one sample per entry (the GPT-5 endpoint samples stochastically at the requested T=0; see E3).

(E2) A third-provider late-cutoff control. As a check that the FPR resistance is not specific to the Anthropic pipeline, we evaluate DeepSeek-V4-Pro—a third-provider model—on a fixed stratified n=300 subsample of the supplement, alongside GPT-5.1 and Sonnet 4.6 re-run on the same entries. This is weaker than the intended control: DeepSeek-V4-Pro *matches* the Anthropic cutoff rather

[4] The control supports the cross-provider reading but does not establish it: DeepSeek-V4-Pro matches rather than exceeds the Anthropic cutoffs, it is a single third-party model, and no recall probe was run on it, so contamination is not separable (§ F.4).

than exceeding it, and the reading below rests on the cross-model *ordering*, not its absolute FPR. On the complete run its FPR is 0.36 at full coverage (two residual UNCERTAIN verdicts; DR 0.87), versus 0.93 for GPT-5.1 and 0.29 for Sonnet 4.6 on the same entries, and its year breakdown (0.29 on 2024 entries, 0.45 on 2025) shows the same graded recency pattern as the cohort. The absolute levels here sit above Tab. 28 (e.g. Sonnet 0.29 vs. 0.120) because this control ran on a later endpoint snapshot and an n=300 subsample; we read the within-run cross-model ordering, not the level (§ 5). A non-Anthropic model thus also resists the post-cutoff FPR rise, consistent with the effect tracking training recency *across providers*. Caveats: a single third-party control; provider-reported cutoff; contamination is not separable for this model (no recall probe run on it).

(E3) Run-to-run variance at the GPT-5 endpoint. We send `gpt-5.1` `temperature=0.0`, but the GPT-5 endpoint samples stochastically regardless of the requested value, so each baseline run is one stochastic draw. Three independent GPT-5.1 runs on a fixed n=150 `dev_public` subsample give F1 $= 0.815 \pm 0.002$, FPR $= 0.528 \pm 0.009$, DR $= 0.965 \pm 0.000$ (mean $\pm$ sample std; $1/150$ label flips). The F1 run-to-run std (0.002) is far below the Sonnet 4.6−GPT-5.1 F1 gap (0.069, $\approx$30$\times$) and the Sonnet 4.6−Opus 4.7 gap (0.016, $\approx$7$\times$), so single-run rankings are stable to sampling noise. Caveat: N=3 runs; n=150 subsample.

Takeaway. The recall probe separates the two readings of a low post-cutoff FPR: Opus 4.7 accepts 84% of valid 2024–2025 papers it *cannot* recall, so its resistance is largely calibration; Sonnet 4.6's gap (87% vs. 50%) is partly abstention: it says UNCERTAIN on a quarter of the papers it cannot recall; GPT-5.1 recalls none and flags 92%. A third-provider control (DeepSeek-V4-Pro) resists alongside them (read the within-run ordering, since this run used a later endpoint snapshot), and GPT-5.1's run-to-run F1 std (0.002) sits ∼7× below even the smallest headline F1 gap (0.016), so single-run rankings are stable to sampling noise. Caveats: title-given recall lower-bounds memorization; n=150–300 per probe.

F.5 THINKING-BUDGET REGIME BOUNDARY

Motivation. The main-table baselines all share a fixed `max_completion_tokens` budget (1024 tokens on the OpenAI/OpenRouter endpoints, including the chain-of-thought DeepSeek-R1 baseline; 256 on the Anthropic native endpoint; see § B.2.4). At this budget, the structured JSON-output contract holds across the entire cohort. Three families of post-snapshot thinking-tier models break this contract—Gemini 3.1 Pro/Flash-Lite, DeepSeek-V4, and Qwen3.5—emitting chain-of-thought scratchpad before the verdict and exhausting the budget mid-reasoning. During the codebase additions for the Q2 2026 frontier cohort we observed concrete failure modes that motivated their exclusion from the main table:

- `google/gemini-3.1-pro-preview`: unbounded thinking tokens overrun the 1024-token `max_completion_tokens` budget, causing JSON parse failures across nearly all entries; `google/gemini-2.5-pro` (which also thinks by default, but mostly completes within the 1024-token budget) is the closest reliable substitute and is included in the main table.
- `google/gemini-3.1-flash-lite-preview`: supports the full thinking range (`minimal`→`high`); same unbounded-thinking risk as gemini-3.1-pro-preview.
- `qwen/qwen3.5-397b-a17b`: ∼40% empty responses (thinking mode consumed the budget before the JSON object closed).
- `qwen/qwen3.5-122b-a10b`: parses successfully but emits 500+ reasoning tokens per reply, inflating cost ∼5× versus the matched non-thinking baseline.
- `deepseek/deepseek-v4-pro` and `deepseek-v4-flash`: thinking models with `reasoning_effort` levels `high`/`xhigh` supported; same budget-exhaustion risk as the V3-R1 chain-of-thought baseline at 1024 tokens.

Smoke-test design. To quantify the regime boundary rather than assert it, we run a stratified n=100 subsample of `dev_public` (proportional across the 11 main hallucination types, ≥ 5 entries per type, fixed seed) on three post-snapshot models under two budget regimes:

Reported metrics. Per (model, regime) cell we report DR, FPR, F1, MCC, parse-failure rate (PF) (fraction of entries whose response is missing or non-JSON), mean and p95 output tokens, and $/entry. Cost is computed at the OpenRouter listed rate as of 2026-05-04.

Results. Tab. 34 reports parse-failure rate, saturation ratio p_{95}/cap, and full classification metrics for all nine (model, regime) cells; Tab. 35 gives the per-type detection-rate breakdown.

Three archetypes of regime-boundary behavior. The smoke test reveals three qualitatively distinct responses to budget headroom:

Table 33: **Smoke-test cohort and budget regimes.** Regime A is a deliberately tight probe at or below the main-table non-thinking budget (1024 tokens on the OpenAI/OpenRouter endpoints; GPT-5.5 is additionally probed at 256 to expose its saturation boundary, Tab. 34); Regime B grants thinking-tier headroom equivalent to $\sim 8\times$ that budget. Parse-failure rate (PF) is the primary diagnostic: a model that exceeds PF ≥ 0.30 under Regime A is excluded from the main table on methodological grounds (the JSON contract does not hold), even if Regime B recovers competitive accuracy.

Model	Tier	Cutoff	Regime A	Regime B
GPT-5.5	non-reasoning	May 2026	256 tok	1024 tok
Gemini 3.1 Pro	thinking	$\geq$Q1 2026	2048 tok (`thinking_budget` 1024)	8192 tok (`thinking_budget` 4096)
DeepSeek-V4-Pro	thinking	$\geq$Q1 2026	4096 tok (`reasoning_effort` low)	8192 tok (`reasoning_effort` high)

Table 34: **Thinking-budget regime boundary smoke test on stratified n=100 subsample of `dev_public`.** PF = parse-failure rate; p_{95}/cap = saturation ratio (1.00 = the 95th-percentile output hits the budget cap). DR / FPR / F1 / MCC computed treating UNCERTAIN and parse-failure entries as label=`VALID`. $ is the OpenRouter list-price cost for the cell at May 2026 rates. Red-shaded saturation cells mark the regime where the cap is biting (p_{95}/cap ≥ 0.95), matching the shaded band of Fig. 7.

Model	Budget	n	PF	p_{95}/cap	DR ↑	FPR ↓	F1 ↑	MCC ↑	mean tok	p_{95} tok	$
GPT-5.5	`256`	100	0.66	1.000	0.100	0.000	0.182	0.180	249	256	$0.87
	`1024`	100	0.12	1.000	0.557	0.000	0.716	0.523	510	1024	$1.65
	`4096`	100	0.01	0.339	0.643	0.000	0.783	0.592	602	1389	$1.93
Gemini 3.1 Pro	`2048 + 1024 reas.`	100	0.00	0.998	0.657	0.033	0.786	0.573	834	2044	$1.34
	`8192 + 4096 reas.`	100	0.01	0.969	0.629	0.033	0.765	0.548	1183	7937	$1.87
	`16384 + 8192 reas.`	100	0.01	0.964	0.643	0.033	0.776	0.560	1688	15793	$2.62
DeepSeek-V4-Pro	`4096, effort=low`	100	0.06	1.000	0.600	0.000	0.750	0.557	1282	4096	$0.16
	`8192, effort=high`	100	0.00	0.366	0.700	0.033	0.817	0.611	1240	3002	$0.16
	`16384, effort=high`	100	0.00	0.211	0.643	0.000	0.783	0.592	1355	3463	$0.17

- **Clean ceiling (GPT-5.5).** Saturation 1.00 → 1.00 → 0.34 across the three regimes; parse-failure rate falls 66% → 12% → 1%; F1 climbs 0.18 → 0.72 → 0.78. The model has an implicit reasoning trace that consumes the entire `max_completion_tokens` budget at 256, and its capability is fully recovered at 4k. This is the textbook regime-boundary behavior.
- **Budget-filling (Gemini 3.1 Pro).** Saturation 0.998 → 0.969 → 0.964 from 2k to 16k; mean output grows monotonically (834 → 1183 → 1688) and the worst-case reasoning trace tracks the cap rather than converging to a fixed length. F1 is essentially flat (0.79 / 0.77 / 0.78) and parse failure stays at $\leq 1\%$, so the cap is not biting in absolute terms. Because its thinking budget was set to half the cap in every cell, output length grows with the cap by design; whether the model limits itself would need a test that raises the cap while keeping the thinking budget fixed. The extra thinking does not help either: F1 stays flat and parse failures stay at or below 1%.
- **Clean recovery (DeepSeek-V4-Pro).** Saturation 1.00 → 0.37 → 0.21; F1 0.75 → 0.82 → 0.78. Recovers cleanly at 8k under `reasoning_effort=high`, and 16k adds nothing. Best-cell F1 (0.82) is competitive with the main-table cohort.

Reading the table. None of the three models would have replaced a main-table baseline in expectation: their best-cell F1 (GPT-5.5: 0.78; Gemini 3.1 Pro: 0.79; DeepSeek-V4-Pro: 0.82) sits in the existing cohort's mid-range, comfortably under the cohort-leading independent F1 (Sonnet 4.6's 0.893). The protocol-budget concern is empirically resolved for two of three: GPT-5.5 and DeepSeek-V4-Pro reach a clean budget ceiling and fit the fixed-budget protocol once we know what budget to allocate. The third (Gemini 3.1 Pro) fills whatever thinking budget it is given: its cost grows with the budget ($1.34 → $2.62 per 100 entries) while F1 stays flat. More generally, a fair fixed-budget benchmark must either (i) allocate the highest budget needed by any one model (inflating cost $\sim 10\times$ across the cohort), or (ii) use a budget at which the JSON contract holds for all listed models and exclude models that cannot answer within it. We choose (ii) and report the boundary explicitly.

Cost and wall-clock. The full nine-cell smoke test cost $10.77 at OpenRouter list prices (May 2026) and consumed ∼303 wall-clock minutes (most of which was DeepSeek-V4-Pro's chain-of-thought latency). Scaling to full `dev_public` (n=1,119) at the same price ratios costs ∼$120, well within a typical per-paper evaluation budget. Wall-clock per cell is ∼10–65 s/entry depending on tier and budget; cells parallelize across providers without contention.

Table 35: **Per-type detection rate across smoke-test cells** (stratified n=100 subsample, ≥ 5 per type). Each cell reports the fraction of hallucinated entries of that type the model labeled `HALLUCINATED`; UNCERTAIN and parse failures count as non-detections. Per-type n is small (≥ 5); 95% binomial CI width is roughly ± 35 pp at n=5, so within-row differences across budgets are noise unless they exceed that band. Useful for relative shape (which types each model finds easiest/hardest at each budget), not for fine absolute comparisons.

Type	Tier	n	GPT-5.5 A	GPT-5.5 B	GPT-5.5 C	Gemini 3.1 Pro A	Gemini 3.1 Pro B	Gemini 3.1 Pro C	DeepSeek-V4-Pro A	DeepSeek-V4-Pro B	DeepSeek-V4-Pro C
`fabricated_doi`	1	5	0.20	0.80	0.80	0.80	0.60	0.60	1.00	1.00	0.80
`nonexistent_venue`	1	5	0.00	1.00	1.00	0.60	0.80	0.60	1.00	1.00	1.00
`placeholder_authors`	1	5	0.00	1.00	1.00	1.00	1.00	1.00	0.80	1.00	1.00
`future_date`	1	5	0.60	1.00	0.80	1.00	1.00	1.00	1.00	1.00	0.80
`chimeric_title`	2	5	0.20	0.40	0.40	1.00	1.00	1.00	0.40	0.60	0.40
`wrong_venue`	2	5	0.00	0.60	1.00	0.40	0.40	0.60	0.80	0.80	0.60
`swapped_authors`	2	5	0.00	0.20	0.60	0.40	0.40	0.40	0.40	0.60	0.40
`preprint_as_pub.`	2	5	0.00	0.40	0.40	0.40	0.40	0.40	0.60	0.60	0.80
`hybrid_fabrication`	2	5	0.20	0.60	0.80	0.80	0.80	0.80	0.60	0.80	0.80
`near_miss_title`	3	5	0.00	0.20	0.40	0.40	0.20	0.40	0.40	0.40	0.40
`plausible_fabrication`	3	5	0.00	0.00	0.20	0.60	0.40	0.40	0.00	0.40	0.20
`merged_citation`	S	5	0.20	0.80	0.80	0.80	0.80	0.80	0.40	0.60	0.60
`partial_author_list`	S	5	0.00	0.00	0.00	0.20	0.40	0.20	0.20	0.00	0.20
`arxiv_version_mismatch`	S	5	0.00	0.80	0.80	0.80	0.60	0.80	0.80	1.00	1.00

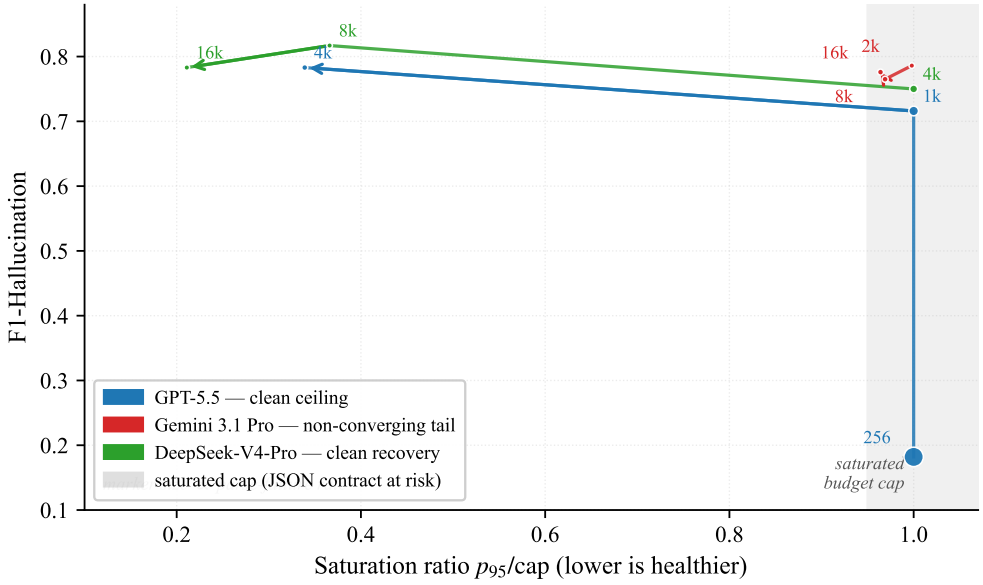

Figure 7: **Thinking-budget regime boundary.** F1 vs. saturation ratio p_{95}/cap for each (model, budget) cell, where p_{95} is the 95th-percentile output token count and cap is the configured `max_completion_tokens`; saturation $\to 1$ means the reasoning trace is hitting the budget ceiling and the structured output is likely truncated. Arrows trace each model's trajectory from low to high budget; marker size is proportional to parse-failure rate; the shaded right band (sat > 0.95) flags the regime where the cap is biting. Three archetypes emerge. GPT-5.5 (blue) escapes the saturated band once the budget exceeds 1k tokens. DeepSeek-V4-Pro (green) *recovers*—saturation drops well below 1.0 once the budget exceeds its typical reasoning-trace length, freeing the JSON contract to complete reliably—and plateaus. Gemini 3.1 Pro (red) stays inside the saturated band at every tested budget, with its thinking budget set to half the cap, so its output grows with the budget by design.

Takeaway. The main cohort is fixed-budget (1024 tokens on the OpenAI/OpenRouter endpoints these smoke-test models use) by construction. Thinking-tier models break the JSON contract at that budget, but two of three (GPT-5.5, DeepSeek-V4-Pro) recover cleanly when the budget bites; one (Gemini 3.1 Pro) uses whatever thinking budget it is given; once outputs are no longer truncated, larger budgets do not raise F1 for any of the three. None would have replaced an existing baseline; HALLMARK will track thinking-tier as a separate cohort once budgets are calibrated.

G BIBTEX-UPDATER: THE CO-DESIGNED REFERENCE TOOL

This section collects everything on `bibtex-updater`, the co-designed reference tool: its implementation (§ G.1), its benchmark evaluation and the co-design-bias controls (§ G.2), and a tool-augmented LLM baseline (§ G.2).

G.1 IMPLEMENTATION

We additionally describe `bibtex-updater`, a co-developed open-source citation verification tool designed for deployment by venues, reviewers, and authors: not to optimize benchmark scores, but to provide reliable, automated checking that integrates into existing publication workflows.

G.1.1 DESIGN GOALS

Three practical requirements drive the design:

1. **Zero human effort.** Verification must be fully automated: no manual review, no prompt engineering, no LLM inference costs. This rules out approaches requiring human-in-the-loop validation or expensive API calls to language models.
2. **Workflow integration.** The tool must integrate into existing pipelines: CI/CD (GitHub Actions), pre-commit hooks, Overleaf builds, and one-off command-line checks. A tool that requires a separate platform or manual invocation will not be adopted.
3. **Graceful degradation.** When APIs are unavailable or rate-limited, the tool should return partial results rather than fail silently. Venues processing hundreds of submissions cannot tolerate flaky infrastructure.

G.1.2 VERIFICATION PIPELINE

`bibtex-updater` implements a multi-stage pipeline that processes each BibTeX entry through increasingly expensive checks:

Pre-API validation (zero cost). Before any network calls, the tool checks for syntactic red flags: DOIs that fail to resolve (HEAD request to `doi.org`, counting only `doi.org`'s own rejection as non-resolution—a publisher-side block or bot-mitigation status reached after a successful redirect is not evidence the DOI is absent, and a versioned arXiv DOI is normalized to its unversioned form first), future publication years, implausible dates (< 1800), and malformed fields. These cheap checks catch Tier 1 hallucinations without API overhead.

Multi-source lookup. The tool queries Crossref, DBLP, and Semantic Scholar using title and first-author search, and resolves arXiv IDs against the arXiv Atom API for preprint-specific consistency checks. Each source returns candidate records that are scored using a weighted combination of fuzzy title matching (70%, token-sort ratio) and author Jaccard similarity (30%). The best-scoring candidate across all sources is selected for field-by-field comparison.

Post-match analysis. Once a candidate is identified, the tool compares DOI, title, authors, year, and venue against the input entry. Venue comparison uses alias-aware matching for 17 major ML/AI venues (e.g., NeurIPS/NIPS, ICML, ICLR, CVPR), so that name variations within that set do not trigger false positives; ISO-4 journal abbreviations outside it (`ACM Trans. Graph.`, `Proc. Natl. Acad. Sci. U.S.A.`) fall below the match threshold and were a substantial false-positive source in the version evaluated here. A dedicated preprint detection stage queries Semantic Scholar to identify entries that claim venue publication when only an arXiv preprint exists.

Status assignment. Each entry receives one status code. Thirteen occur on `dev_public`: *verified* (442), *unconfirmed* (147), *swapped_authors* (146), *partial_match* (85), *venue_mismatch* (66), *not_found* (52), *doi_not_found* (46), *doi_mismatch* (36), *future_date* (30), *arxiv_id_mismatch* (18), *year_mismatch* (17), *title_mismatch* (14) and *given_name_substitution* (13). The 147 *unconfirmed* verdicts are where most of the tool's abstention sits. The HALLMARK wrapper maps these to binary labels and confidence scores for benchmark evaluation.

G.1.3 DEPLOYMENT MODES

The tool supports three deployment scenarios: CI/CD integration (`--strict` flag exits with nonzero code on detection, gating submission workflows), pre-commit hooks (validates `.bib` files on every commit), and batch processing (concurrent workers with rate limiting for venue-scale use). For venue-scale deployment, we recommend a two-stage workflow: run `bibtex-updater` to flag suspicious citations, then manual review of high-confidence flags (confidence ≥ 0.7); at a 2% base rate expect roughly one true hallucination per six flags, and read that as the optimistic end—at

Table 36: **Results for `bibtex-updater` v1.2.0 (co-designed alongside this benchmark) on `dev_public`.** Reported separately from independent tools in Tab. 1 due to co-design bias concerns. The tool abstains when no backing record is found; its DR/FPR/F1 score abstentions as committed-VALID over the full split (the headline convention), while Cov. reports the 0.87 fraction it commits to. Its selective and conservative/aggressive split is in § E.2.

Tool	DR ↑	FPR ↓	F1 ↑	MCC ↑	TW-F1 ↑	ECE ↓	Cov.
bibtex-updater	0.820	0.051	0.880	0.768	0.892	0.395	0.86

1% it is one per twelve and at 0.5% one per twenty-two, and the prevalence we measured on real submissions was indistinguishable from zero (§ E.1 and § 5). At NeurIPS scale (approximately 10,000 submissions, approximately 50 references each) that is half a million references. At the throughput we measure with keys, credentials and a 20-per-minute pace (2.4 s per entry) eight workers need roughly 42 hours; at the wrapper's older rate limit (5.7 s per entry) roughly 99. The tool requires no GPU or LLM API keys and is MIT-licensed; code is publicly available at `https://anonymous.4open.science/r/hallmark/` (anonymized for review).

G.2 Evaluation and co-design bias

`bibtex-updater` (Anonymous, 2025) is a multi-database cross-referencing tool co-developed alongside this benchmark. Its results appear in Tab. 1 under the "Co-designed (reference upper bound)" section, separated from independent tools by a rule. We include it in the main table to give a complete picture of the recall–precision spectrum: `bibtex-updater` anchors the precision end by flagging conservatively (DR 0.820, FPR 0.051), not by topping detection rate, and it abstains on the ∼12–13% of entries it cannot verify against a backing record (coverage 0.87 dev / 0.88 test), an abstention that recovers recall; the low FPR comes from conservative matching, not from abstaining. The explicit upper-bound label and this appendix keep the comparison transparent. The co-design relationship means the benchmark taxonomy and sub-test structure were informed by the tool's detection capabilities, creating a potential circularity that could inflate its apparent performance; readers should treat its numbers as an informative upper bound, not a fair head-to-head comparison with independent tools.

Per-type breakdown for bibtex-updater. Tab. 37 shows per-type detection rates for bibtex-updater. It detects perfectly (a per-type detection rate of 1.000; parenthesized values throughout this passage) on `fabricated_doi`, `placeholder_authors`, `future_date`, `chimeric_title`, `hybrid_fabrication`, `plausible_fabrication`, and `merged_citation`, and stays strong on `swapped_authors` (0.985) and `near_miss_title` (0.923). The weakest types are precisely those where the tool abstains rather than guesses—`partial_author_list` (0.219), `nonexistent_venue` (0.667), `preprint_as_pub.` (0.677), `wrong_venue` (0.681), and `arxiv_version_mismatch` (0.714)—where a could-not-verify verdict is scored as VALID in this forced-binary detection rate (the abstention behavior of § E.2). These cells are scored from the tool's per-entry verdicts on the released labels, the same verdicts behind the aggregate in Tab. 36.
Readers should interpret `bibtex-updater`'s results with caution: (1) its multi-database verification strategy was developed concurrently with the benchmark design; (2) the sub-test structure mirrors the tool's internal verification pipeline; (3) the pre-screening layer was designed to complement its known blind spots. We encourage independent tool developers to submit evaluations via HALLMARK's contribution framework to establish unbiased baselines.
One observation bounds the circularity concern from the other side. The ground-truth audit recovered 52 real papers (27 `dev_public`, 25 `test_public`) that earlier labeling had wrongly marked HALLUCINATED. On 40 of those 52—19 of 27 on `dev_public` and 21 of 25 on `test_public`—`bibtex-updater`'s database cross-check returned VALID, diverging from the then-current (buggy) labels rather than tracking them; it agreed with the buggy label on the remaining 12. A tool whose apparent accuracy were merely an artifact of co-design would mirror the benchmark's labels, including their errors, so a majority divergence is evidence that its detections rest on external database evidence rather than on the label distribution. It is evidence and not proof: 12 entries went the other way. The tool abstains on entries it cannot back with a record, reporting `dev_public` DR 0.820 / FPR 0.051 / F1 0.880 (Tab. 36), and the relabel removed an artifact that had flattered the more aggressive LLM verifiers more than the rule-based tool (§ 4.3).

Cross-domain probe. The released `test_crossdomain` split (500 entries: 200 valid / 300 hallucinated; composition in § A.3) evaluates `bibtex-updater` v1.2.0 outside the ML-venue regime

Table 37: **Per-type detection rates for `bibtex-updater` (v1.2.0) on `dev_public`.** Forced-binary detection rate on the released labels, scored from the tool's per-entry verdicts. Types on which the tool abstains (could-not-verify) score lower, since abstentions count as VALID in the detection rate (§ E.2).

Tier	Type	bibtex-updater
1	fabricated_doi	1.000
	nonexistent_venue	0.667
	placeholder_authors	1.000
	future_date	1.000
2	chimeric_title	1.000
	wrong_venue	0.681
	swapped_authors	0.985
	preprint_as_pub.	0.677
	hybrid_fabrication	1.000
3	near_miss_title	0.923
	plausible_fabrication	1.000
Stress	merged_citation	1.000
	arxiv_version_mismatch	0.714
	partial_author_list	0.219

Table 38: **`bibtex-updater` v1.2.0 in and out of regime.** `dev_public` (ML venues, 2021–2023) vs. the released `test_crossdomain` split vs. a recency-matched, canonically-resolved biomedical rebuild (`matched`: 152 valid entries from 2021–2023, each field resolved to its DOI's CrossRef record). Abstentions scored committed-VALID (the headline convention); Cov. is the committed fraction over all entries. The released split's FPR (red) is inflated by post-cutoff recency and scrape-metadata noise; canonical resolution returns the FPR to its in-domain level (green) while coverage falls to 0.582 (the tool abstains on 71% of the valid biomedical entries).

Split	DR ↑	FPR ↓	F1 ↑	Cov.
dev_public	0.820	0.051	0.880	0.86
test_crossdomain	0.890	0.375	0.832	0.736
test_crossdomain_matched	0.777	0.112	0.847	0.582

the main splits sample. Detection transfers: DR reaches 0.890, against 0.865 on `dev_public`. The released split reports FPR 0.375 (Tab. 38), but two confounds inflate that number. First, the split's valid biomedical entries are all dated 2026, post-cutoff for every LLM (the recency confound we take up below); `bibtex-updater` queries live databases and carries no cutoff, so this does not affect its own verdicts. Second, and specific to the rule-based tool, our biomedical scrape carried metadata noise—Vancouver-style author initials, epub-versus-print year ambiguity, preprint venue strings—that its exact-match verification reads as contradictions and flags, manufacturing false positives that reflect the citation record rather than the tool's judgment. A recency-matched, canonically-resolved rebuild removes both: 152 valid biomedical entries from 2021–2023, each field replaced by the CrossRef record its DOI resolves to (canonical author, year, venue), plus the same three hallucination tiers. On that clean split `bibtex-updater`'s headline FPR is 0.112, about twice its in-domain 0.051, though it commits to only 29% of the valid biomedical entries, abstaining on the other 71% (against 13% in-domain). The canonical metadata converts the released split's mis-flags into honest abstentions: out of domain the tool abstains rather than over-flags, because its ML-tuned assumptions (full author names, one canonical year, exact venue strings) cannot confirm biomedical records against its backing databases. The cost of leaving the regime is coverage, not precision (Tab. 38).

Recovering coverage with a Stage-2 diagnoser. `bibtex-updater`'s out-of-domain abstention is a coverage gap that a second stage can fill. We run the paper's two-stage cascade on the matched split: Stage 1 `bibtex-updater` decides the 263 entries it can confirm or refute and defers the 189 it cannot to Stage 2, an agentic Sonnet 4.6 diagnoser (up to five tool calls per entry).[5] Stage 2

[5]Stage 1 is served from the persisted standalone verdicts; re-running `bibtex-check` live hangs on arXiv DOI HEAD checks against the fabricated future-dated DOIs, so the live cascade is impractical here. Stage 2 is a dated OpenRouter snapshot (§ B.2.4).

Table 39: **Domain × recency decomposition of LLM false positives.** Each cell is FPR on valid entries. Columns cross domain (in = ML `dev_public`; out = biomedical) with recency (pre-cutoff 2021–2023; post-cutoff: in-domain = the 2024–25 temporal supplement, out-domain = the released 2026 biomedical entries). Δ_{dom} is the pure domain effect (out-pre − in-pre, recency held fixed). Two readings: (i) the out-domain post-cutoff column is ≈1.0 for every model, reflecting the post-cutoff date heuristic (89% of those false positives cite the future date) rather than domain transfer; (ii) once recency is matched, Δ_{dom} is negative or small for calibrated verifiers and large only for the recall-aggressive Qwen3-235B. Sorted by out-domain pre-cutoff FPR; DeepSeek-R1 omitted (see text). §Sonnet 4.6 and Opus 4.7 use the May collection in the in-domain columns, not the September re-run of Tab. 1; their out-domain runs are from July.

	pre-cutoff			post-cutoff	
Model	in-dom.	out-dom.	Δ_{dom}	in-dom.	out-dom.
Gemini 2.5 Flash	0.100	0.013	−0.09	0.595	0.991
Gemini 2.5 Pro	0.050	0.020	−0.03	0.250	1.000
Llama 4 Maverick	0.146	0.059	−0.09	0.763	1.000
Claude Sonnet 4.6§	0.127	0.099	−0.03	0.120	0.957
GPT-5.1	0.411	0.105	−0.31	0.759	1.000
GPT-5.4	0.228	0.112	−0.12	0.413	1.000
Claude Opus 4.7§	0.072	0.125	+0.05	0.073	0.948
Mistral Large	0.250	0.178	−0.07	0.793	1.000
DeepSeek-V3.2	0.702	0.408	−0.29	0.759	1.000
Qwen3-VL-235B	0.551	0.553	+0.00	0.887	1.000
Qwen3-235B	0.533	0.974	+0.44	0.809	1.000

recovers the coverage: the cascade commits on all 152 valid biomedical entries (valid-entry coverage $0.29 \rightarrow 1.00$) at FPR 0.189 and DR 0.975 (F1 0.941; the aggressive stance is nearly identical at FPR 0.211, since Stage 2 leaves almost nothing uncertain). Closing the coverage gap costs little precision—headline FPR 0.112 with `bibtex-updater` abstaining, 0.189 with the cascade committing on everything—and stays well below the recall-aggressive models' over-flagging (Qwen3-235B 0.974): out of domain, the abstention is recoverable.

The LLM cohort: domain versus recency. The zero-shot LLM verifiers of Tab. 1 were unevaluated cross-domain; we run eleven of them on both the released 2026 biomedical entries and the recency-matched rebuild, which turns the released split's confound into a clean 2 × 2 over domain and recency (Tab. 39).[6] The released split's high biomedical FPR is a post-cutoff artifact: 89% of the biomedical false positives there give the future date as the reason ("2026 is in the future," "beyond my training cutoff"), and on the recency-matched entries—same domain, pre-cutoff—that share falls to 2%, with the remaining false positives turning into genuine recall failures (an unresolvable DOI, an unfamiliar venue). The residual pure-domain effect (matched FPR minus `dev_public` FPR, recency held fixed) is model-dependent: negative or small for the calibrated verifiers (GPT-5.1 −0.31, GPT-5.4 −0.12, Sonnet 4.6 −0.03, Opus 4.7 +0.05, Gemini 2.5 Pro −0.03, Llama 4 Maverick −0.09), and large only for a recall-aggressive model (Qwen3-235B +0.44), which flags legitimate bioRxiv preprints cited in article form as fabrications. Biomedical citations are not inherently harder to verify: the cross-domain FPR rise was recency, and what domain adds is small except where a model already over-flags in its own regime.

Provider content filtering on biomedical citations. Both Anthropic models we query through OpenRouter (Claude Opus 4.7 and Sonnet 4.6) return an empty, content-filtered response (`finish_reason: content_filter`) on the same 5 of 152 valid biomedical entries (3.3%) in the recency-matched split: all virology and immunology papers on SARS-CoV-2 inhibitor resistance, ACE2 receptor-binding and antibody-escape mutations, and HIV envelope neutralization, i.e. dual-use–adjacent titles the provider filter blocks. The refusal is deterministic: it reproduces at completion budgets from 512 to 4096 tokens, so it is not output truncation, and a trivial control prompt to the same endpoint returns normally, so it is not an outage or a credit fault. These entries fall to the verifier's error path and score as UNCERTAIN, i.e. committed-VALID under our convention, so they do not inflate false-positive rates. They do depress detection: the same error path catches

[6] DeepSeek-R1 is excluded: 432 of its 500 cross-domain predictions are `[Error fallback]` API failures (86%), so no metric on that split measures the model.

Table 40: **Tool-augmented GPT-5.1 vs. standalone components on `dev_public`.** The augmented model improves precision and calibration over GPT-5.1 alone but does not match bibtex-updater's detection rate.

Baseline	DR ↑	FPR ↓	F1 ↑	MCC ↑	TW-F1 ↑	ECE ↓
GPT-5.1 (standalone)	0.837	0.411	0.766	0.442	0.822	0.190
bibtex-updater (standalone, v1.2.0)	0.820	0.051	0.880	0.768	0.892	0.395
GPT-5.1 + bibtex-updater	0.843	0.144	0.856	0.698	0.872	**0.078**

10 hallucinated entries for Opus and 9 for Sonnet, which then score as missed detections, so the cost is one of availability for the FPR column and of accuracy for the DR column. The released prediction dumps record every one of these records as `[Error fallback] Parse error: None`, indistinguishable there from a generic API failure; the `content_filter` finish reason was observed at run time and is not recoverable from the release. The filter never fires on the ML-venue core, so the failure mode is invisible in-domain and surfaces only out of domain: a deployment consideration specific to safety-filtered providers, orthogonal to the verdict-quality question the rest of this section studies.

Takeaway. The cross-domain FPR rise is mostly a recency artifact, not a domain effect. For the LLMs, 89% of the released biomedical false positives are post-cutoff date heuristics that fall to 2% once the entries are pre-cutoff; the residual domain effect is negative or small for calibrated verifiers and large only for a model that already over-flags in its own regime. For `bibtex-updater`, canonical metadata shows the released 0.375 was inflated by scrape noise: its true out-of-domain behavior is a coverage drop—it abstains on 71% of valid biomedical citations—at an in-domain-level FPR (0.112). A Stage-2 Sonnet diagnoser recovers that coverage ($0.29 \rightarrow 1.00$ on valid entries) at FPR 0.189: the out-of-domain abstention is a recoverable gap. Precision is a property of the tool–regime pair, and the pair that matters is domain *and* recency together.

Tool-augmented LLM baseline. To test whether combining LLM reasoning with API-backed verification is greater than the sum of its parts, we augment GPT-5.1 with `bibtex-updater`'s structured output. For each entry, we first run `bibtex-check` to obtain the verification status, mismatched fields, confidence score, and APIs consulted, then inject this evidence into an augmented prompt that instructs the model to use the tool findings as evidence while applying its own judgment. Tab. 40 compares the augmented model against its components.

The augmented model improves over standalone GPT-5.1 on every headline metric: detection rate rises slightly (84.3% vs. 83.7%) while FPR drops sharply (0.144 vs. 0.411) and ECE improves to 0.078, the best calibration of any baseline. However, it still falls short of `bibtex-updater`'s standalone detection rate (84.3% vs. 86.5%, where the tool abstains rather than guessing on unbacked entries).

Tab. 41 reveals why. On the two types where bibtex-updater excels and GPT-5.1 struggles—`swapped_authors` and `near_miss_title`—the augmented model tracks the LLM rather than the tool: `swapped_authors` reaches 47.8% against bibtex-updater's 98.5% (standalone GPT-5.1: 44.8%), and `near_miss_title` falls from the LLM's own 57.7% to 50.0% against the tool's 92.3%. The LLM overrides tool-detected metadata mismatches, treating them as potential API artifacts rather than genuine hallucination signals. The one type where evidence injection buys anything is `chimeric_title`, which rises to 97.9% from standalone GPT-5.1's 89.4%: the tool evidence helps the LLM *confirm* an existing suspicion rather than revise a judgment. On three types—`fabricated_doi`, `near_miss_title` and `plausible_fabrication`—the augmented model lands below both of its components at once.

These results demonstrate that naïve evidence injection—presenting tool output as LLM context—is insufficient to realize the complementarity between parametric and retrieval-based verification. The LLM treats tool evidence as *advisory* rather than *authoritative*, defaulting to its own judgment when conflicts arise. More structured integration strategies—such as forcing acceptance of tool-detected field mismatches, weighted ensembling, or agentic tool-use where the LLM iteratively queries APIs—may better combine the complementary strengths.

H Robustness ablations

We report four ablations that probe whether the precision-ceiling finding is an artifact of a single design choice—a prompt, a threshold, an input field, or a single annotator—alongside the selective-prediction analysis of § E.2. We read all four as *robustness evidence for the ranking*, not as tuning for best numbers: the tool ranking is what transfers across the regimes these ablations vary (§ 4.3)—it

Table 41: **Per-type detection rates for tool-augmented GPT-5.1 vs. standalone components.** The augmented model fails to transfer `bibtex-updater`'s strength on metadata-based types. All three columns are scored on the released labels; the **GPT-5.1** column reproduces the `per_type_metrics` block of `llm_openai_dev_public.json`. Red marks the three cells where evidence injection falls below both standalone components.

Tier	Type	GPT-5.1	bibtex-updater	Augmented
1	fabricated_doi	1.000	1.000	0.923
	nonexistent_venue	1.000	0.667	0.795
	placeholder_authors	1.000	1.000	1.000
	future_date	1.000	1.000	1.000
2	chimeric_title	0.894	1.000	0.979
	wrong_venue	0.851	0.681	0.681
	swapped_authors	0.448	0.985	0.478
	preprint_as_pub.	0.742	0.677	0.677
	hybrid_fabrication	1.000	1.000	1.000
3	near_miss_title	0.577	0.923	0.500
	plausible_fabrication	0.934	1.000	0.882

does *not* survive a change in what counts as a hallucination, which § H.1 measures separately—so we test its invariance to phrasing (§ H.2), operating point (§ H.4), input format (§ H.5), and rater (§ H.6).

H.1 What counts as a hallucination changes the ranking

Four of the fourteen types label entries whose cited work exists. `wrong_venue` assigns a real paper to the wrong conference, `preprint_as_published` cites a real preprint as though it appeared at a venue, `partial_author_list` truncates a real author list without marking it, and `arxiv_version_mismatch` cites a real preprint with the wrong venue and a shifted year. In each case the reference points at a work a reader can find; what is wrong is the description. The other ten types cover works that do not exist. A verifier that flags all fourteen is doing two jobs, and the benchmark scores them as one.

The distinction is not marginal. On `test_public` the four modes are 139 of 519 hallucinated entries (26.8%), and on `dev_public` 158 of 606 (26.1%). They are not confined to `stress_test`, and the stress split is 61.5% real-paper modes in its own right.

We re-score all 20 tool configurations three ways (Tab. 42): as shipped; with the four modes folded out of the evaluation set entirely; and with them scored as false accusations, so a flag on one counts against the tool. The ranking is stable to neither. Folding them out moves 12 of 20 configurations in rank; scoring them as false accusations moves 16 of 20 and changes the leader. The cascade of Tab. 1 leads as shipped at MCC 0.941 (0.938 conservative) and falls to 0.668 (0.665) as false accusations; `bibtex-updater` falls less, from 0.738 to 0.688, and takes first place under that reading. Folded out, the cascade keeps its lead (0.936 against 0.845). The reversal has a plain cause, visible in the tools' detection rates on the four modes alone: the cascade catches nearly all of them (0.993) while `bibtex-updater` catches 0.597 and DOI-only 0.065. More of the cascade's headline credit comes from flagging papers that exist.

A second control bounds how much verification those detections require. A classifier with access only to surface features—which fields are present, how long the strings are, how many authors there are, whether the venue string says arXiv, the entry type, the year—and no network access and no semantics reaches DR 0.468 at FPR 0.045 on `dev_public` and DR 0.388 at FPR 0.096 on `test_public`, against a majority-class baseline of 0.000. Of the four modes' positives 80% are perturbations of real entries, while 95% of the valid entries they are scored against were scraped from real bibliographies, so roughly 40% of the separation is a property of how the entries were made rather than of what they claim.

We report this rather than relabel. Which of the four modes a deployed verifier should flag depends on what the verifier is for: a venue chair screening for fabricated citations wants the ten fabrication modes, while a copy-editing pass wants all fourteen. The benchmark as released serves the second reading, and Tab. 1 should be read as a ranking under it. Readers with the first use in mind should take the folded-out column, where the cascade still leads `bibtex-updater` (0.936 against 0.845). Public venue guidance does not agree on this line. ICML 2026's area-chair instructions count a reference as hallucinated only when no document with a title anywhere near it can be found, and exclude a real

venue that is incorrect for the reference, a swapped author list, and a few missing authors (ICML, 2026): `wrong_venue`, `preprint_as_published` and `arxiv_version_mismatch` fall on the excluded side, and `partial_author_list` does when the truncation is small. ICLR 2026's review retrospective instead describes flagged references as ones that "referred to documents that didn't exist and/or contained egregiously incorrect bibliographic information" (ICLR, 2026), merging the two classes, and ACL 2026 lists hallucinated references as grounds for desk rejection without defining them (ACL, 2026). We therefore report both readings rather than choosing one.

Table 42: **What counts as a hallucination changes the ranking (`test_public`, 20 configurations).** Real-mode DR is the detection rate on the four types whose cited work exists (`wrong_venue`, `preprint_as_published`, `partial_author_list`, `arxiv_version_mismatch`); MCC is reported as shipped, with those four types folded out of the evaluation set, and with them scored as false accusations. Rows are named by their released result file; the cascade rows are the Sonnet 4.6 cascade on the shared Stage 1, and `bibtexupdater` is the rescored run.

Configuration	Real-mode DR	MCC as shipped	MCC folded out	MCC as false accusations
`cascade_db_diagnosis`	0.992	0.938	0.932	0.665
`cascade_db_diagnosis_aggressive`	0.993	0.941	0.936	0.668
`llm_openrouter_claude_sonnet_4_6`	0.863	0.743	0.767	0.569
`bibtexupdater`	0.597	0.738	0.845	0.688
`llm_agentic_btu_sonnet_4_6`	0.993	0.721	0.701	0.512
`llm_openrouter_claude_opus_4_7`	0.791	0.700	0.748	0.573
`llm_agentic_btu_openai`	0.921	0.661	0.668	0.500
`llm_tool_augmented`	0.755	0.601	0.647	0.497
`llm_openai_gpt54`	0.655	0.544	0.603	0.472
`llm_openrouter_deepseek_r1`	0.725	0.489	0.530	0.407
`llm_agentic_openai`	0.899	0.464	0.479	0.366
`llm_openrouter_llama_4_maverick`	0.450	0.451	0.530	0.444
`llm_openrouter_gemini_pro`	0.252	0.427	0.517	0.459
`llm_openrouter_gemini_flash`	0.324	0.404	0.482	0.414
`llm_openai`	0.799	0.397	0.422	0.322
`llm_openrouter_mistral`	0.475	0.394	0.484	0.424
`llm_openrouter_qwen_max`	0.914	0.372	0.374	0.276
`llm_openrouter_qwen`	0.899	0.355	0.356	0.261
`llm_openrouter_deepseek_v3`	0.863	0.244	0.272	0.218
`doi_only`	0.065	0.211	0.273	0.274

The temporal-mechanism probes (the recall and late-cutoff controls, cutoff-aware prompting, and the thinking-budget boundary) live with failure mode (iii) in § F; the bootstrap CIs are in § B.1. The prompt-variant, field-leave-one-out (LOO), and rater runs use a fresh dated OpenRouter snapshot (2026-05-31) that does not reproduce the main-run absolute aggregates; we therefore read each as a *within-run* difference, robust to endpoint drift (§ 5).

H.2 Prompt-sensitivity sweep

We sweep four prompt variants—`default` (the paper's prompt), `notaxo` (taxonomy removed), `uncertain` (abstention explicitly encouraged), and `terse` (compressed instructions)—over four models on a stratified n=150 `dev_public` sample (81 hallucinated / 69 valid), at temperature 0 and seed 42 (Tab. 43); GPT-5.1 runs through its OpenAI-direct endpoint, the other three through OpenRouter.

The finding is two-part. First, *the model ranking is prompt-invariant*: mean pairwise Spearman $\rho = 0.90$ for F1, DR, and FPR across the four variants (Sonnet 4.6 > GPT-5.1 > DeepSeek-V3.2 > Gemini 2.5 Flash on F1 at `default`), and the sole departure from $\rho = 1.0$ is a 0.001 F1 near-tie between GPT-5.1 (0.794) and Sonnet 4.6 (0.793) under the `uncertain` variant: a tie, not a reordering. This is the result that defends the single-prompt design of § 4.2. Second, *the absolute FPR is wording-sensitive*: GPT-5.1 shows the largest swing: FPR 0.580 at `default` down to 0.212 under `terse` (-36.8 pp), with its UNCERTAIN rate reaching 19.3% under `uncertain` (from 0%). The abstention-encouraging `uncertain` variant drops Sonnet 4.6's FPR from 0.121 to 0.015 (-10.6 pp) and DeepSeek-V3.2's from 0.899 to 0.600 (-29.9 pp), while Gemini 2.5 Flash moves only -4.3 pp. Sonnet's UNCERTAIN rate rises from 2.0% (`notaxo`/`terse`) to 8.7% (`uncertain`), a $\sim$7 pp coverage swing from wording alone. Pooled over all models and non-default variants, the verdict-flip rate against the default is 17.4% (all entries) and 13.6% when UNCERTAIN-involving flips are excluded, with GPT-5.1 the most prompt-sensitive model (25.8% mean flip).

The practical consequence is a scope bound on the post-cutoff FPR *magnitude*: because a single wording change can move FPR by 10–37 pp, we treat the absolute post-cutoff false-alarm rate (§ 4.4)

Table 43: **Prompt-sensitivity sweep** (n=150 `dev_public`; temperature 0; seed 42; OpenRouter models on a fresh 2026-05-31 snapshot, GPT-5.1 via the OpenAI-direct endpoint). DR, FPR, F1, ECE, and UNCERTAIN-rate per model and prompt variant. Model F1/DR ranking is near-invariant to the variant (mean pairwise Spearman $\rho = 0.90$); the absolute FPR is not (the `uncertain` and `terse` variants drop it sharply).

Model	Variant	DR ↑	FPR ↓	F1 ↑	ECE ↓	UNC.
Sonnet 4.6	default	0.909	0.121	0.903	0.087	4.7%
	notaxo	0.744	0.101	0.811	0.098	2.0%
	uncertain	0.667	0.015	0.793	0.083	8.7%
	terse	0.759	0.074	0.833	0.113	2.0%
GPT-5.1	default	0.914	0.580	0.759	0.233	0.0%
	notaxo	0.802	0.420	0.743	0.225	0.0%
	uncertain	0.800	0.250	0.794	0.145	19.3%
	terse	0.720	0.212	0.755	0.170	6.0%
DeepSeek-V3.2	default	0.975	0.899	0.712	0.372	0.0%
	notaxo	0.938	0.768	0.724	0.335	0.0%
	uncertain	0.872	0.600	0.735	0.282	4.7%
	terse	0.889	0.667	0.724	0.315	0.0%
Gemini 2.5 Flash	default	0.469	0.130	0.594	0.316	0.0%
	notaxo	0.370	0.087	0.513	0.324	0.0%
	uncertain	0.383	0.087	0.525	0.314	0.0%
	terse	0.321	0.101	0.456	0.414	0.0%
Mean pairwise Spearman ρ		F1/DR: 0.90		FPR: 0.90		

as prompt-conditional and rest the temporal claim on the cross-regime *ranking*, which this sweep shows is prompt-invariant.

Takeaway. Prompt wording moves the absolute numbers and leaves the ranking alone: a single variant shifts FPR by 10–37 pp (GPT-5.1: 0.580 → 0.212 under `terse`) and Sonnet's coverage by ∼7 pp, whereas the model ordering holds at mean pairwise Spearman $\rho = 0.90$, with the sole departure a 0.001 F1 near-tie. This is why the temporal claims rest on cross-regime rankings and every absolute post-cutoff FPR reads as prompt-conditional (§ 4.4).

H.3 GEPA PROMPT OPTIMIZATION

The robustness check above bounds how far wording moves the metrics; it does not say how far a *deliberate* search through wording can move them. That question is sharpened by failure mode (iii): the cutoff experiments show post-cutoff over-flagging (§ 4.4), the cutoff-aware check shows part of it answers to instructions (§ F.2), and the sweep prices wording high enough to cover the FPR gap between a small model and the frontier (§ H.2). If a prompt is worth that much, the precision gap we report might be a prompting effect rather than a capability difference. We test that directly, and the answer is no.

We optimize the paper's verification prompt with Genetic-Pareto (GEPA), a reflective evolutionary search that mutates a prompt from natural-language feedback on its own failures (Agrawal et al., 2026). The seed is the paper's default prompt (§ B.2.3) and the model is Claude Haiku 4.5. The objective is balanced binary accuracy on a 200-entry `dev_public` validation set and under-flagging is penalized as hard as over-flagging; UNCERTAIN scores 0. **Optimization touched `dev_public` only; every number below is `test_public`**, which the optimizer never saw.

Optimization works, but does not close the gap. Claude Opus 4.7 on the untuned default prompt catches far more hallucinations than the optimized small model at a similar false-positive rate (DR 0.900 vs. 0.746, FPR 0.205 vs. 0.224) on the same split (Tab. 9). *Prompt optimization does not move a small model to the frontier's operating point.* We state the claim no more broadly than that: only Haiku was optimized, and it is compared against frontier models running the shared default prompt, so this is an optimized-vs-default comparison.

A note on protocol, since prompt optimizers leak by default. The reflection step sees per-entry feedback on failures, which is an easy channel for the optimizer to write dataset specifics into the prompt. We therefore constrain reflection to describe failure patterns in general terms and forbid it from copying author names, titles, citation keys, or split statistics. The optimized prompt is released as `results/gepa_haiku/guarded/best_prompt.txt`.

Table 44: **GEPA prompt optimization on Claude Haiku 4.5.** Both rows are the same model under the same decoding settings; only the prompt differs. Bal. is balanced accuracy, $(\text{DR} + 1 - \text{FPR})/2$, on `test_public`; Val. objective is the same metric on the 200-entry `dev_public` validation set the optimizer tuned on. Abstentions score 0 in both.

	Val. objective	**Held out: `test_public`** (n=831)		
Prompt	`dev_public` (n=200)	**DR** ↑	**FPR** ↓	**Bal.** ↑
Seed (paper default)	0.700	0.871	0.468	0.702
GEPA-optimized	0.800	0.746	0.224	0.761
Δ	+0.100	−0.125	−0.244	+0.059

H.4 Threshold and aggregation ablation

We re-score the stored per-entry confidences on `dev_public` (no API calls) to test two operating-point choices: the decision threshold and the field-aggregation rule (Tab. 45).

Threshold. Confidences are effectively quantized, so threshold tuning buys almost nothing: the gap between the default 0.5 threshold and the best-F1 threshold is below 0.35 pp for seven of eight tools (GPT-5.4 0.34 pp, Opus 4.7 0.20 pp, Sonnet 4.6 0.09 pp, the rest ≤0.07 pp). The lone exception is Gemini 2.5 Flash (8.22 pp), and it buys F1 only by trading FPR up to 0.343. AUROC orders the verifiers as expected (Sonnet 0.927, Opus 0.903, GPT-5.4 0.834, then the recall-aggressive open-weight models 0.61–0.74). For the seven quantized-confidence tools the fixed-0.5 operating point of Tab. 1 is therefore near-optimal—best-F1 is chosen in-sample on `dev_public`, so this lower-bounds the true headroom—while Gemini 2.5 Flash is the exception noted above. The ranking is threshold-robust regardless, in contrast to peer benchmarks that hard-code an unjustified match threshold.

Aggregation. The benchmark's detector flags an entry when *any one* of the cross-database sub-tests fails (the "any-miss" rule—the benchmark's internal detector, distinct from the agentic harness's *any-no-match* rule in § D). Sweeping stricter quorums trades DR for FPR monotonically: on the structured sub-tests, any-miss gives DR 0.950 / FPR 0.000, while requiring two, three, or all four misses reduces DR to 0.269, 0.048, 0.002; on the field-level resolver the same sweep drops DR from 0.759 (any-miss) to 0.128 (unanimous). A noisy-voter ensemble makes the same point with real verifiers: treating the twelve zero-shot LLMs with stored per-entry predictions (Tab. 11) as twelve independent voters, flagging whenever any single voter flags reproduces the any-no-match profile (DR 0.995 / FPR 0.881), simple majority ($\geq 7/12$) gives DR 0.779 / FPR 0.191 / F1 0.803, and supermajority and unanimity buy lower FPR at steep DR cost. The benchmark deploys the any-miss rule (F1 0.975); requiring cross-database agreement instead nudges F1 to 0.985 at a small false-positive cost (FPR 0.000 → 0.035), so any-miss sits within 1 pp of the family's F1 maximum while holding FPR at zero.

H.5 Input-format and field leave-one-out

We test whether detection rests on a surface format tell or on field content by varying the input representation (full BibTeX vs. structured fields) and dropping one field at a time (title / authors / venue / year / DOI), on the same n=150 stratified sample with two models (Tab. 46). This is the dynamic counterpart to the static format-tell audit (§ C.5.4).

Two results. *Title carries most of the signal.* For Gemini 2.5 Flash, dropping the title raises FPR by +35.5 pp over the structured baseline (0.145 → 0.500) and dropping authors by +18.8 pp, while venue, year, and DOI move FPR little; DeepSeek-V3.2—already near-saturated at FPR 0.81 structured—shows the same direction at smaller magnitude (+12.9 pp for title). *Format matters less than content*: switching between full BibTeX and structured fields shifts DR/FPR by at most 7.4/14.5 pp. Detection therefore leans on title and author semantics rather than a surface artifact, reinforcing the construct-validity argument that the result is not driven by a single co-designed field or by formatting (§ G.2 and § C.5.4).

H.6 Multi-rater reliability proxy

This is an automated multi-rater reliability proxy, not human inter-annotator agreement. Three independent LLM raters (Sonnet 4.6, DeepSeek-V3.2, Gemini 2.5 Pro) label 132 blinded entries—80 real-world-incident hallucinations and the 52 relabel-recovered real papers—and we report their agreement (UNCERTAIN mapped to HALLUCINATED, binary; Tab. 47). It does not substitute for

Table 45: **Threshold and aggregation ablation on `dev_public`** (offline re-score of stored confidences; no API calls). *Top:* per-tool AUROC, default-0.5 F1, and the F1 gap to the best threshold—near zero for all but Gemini 2.5 Flash, confirming the fixed operating point is near-optimal. *Bottom:* the field-aggregation sweep; the benchmark uses the any-miss rule, which sits within 1 pp of the family's F1 maximum while holding FPR at zero, and stricter quorums trade DR for FPR monotonically.

(a) Threshold: default 0.5 vs. best-F1 (8 tools)

Tool	AUROC	F1@0.5	best-F1	gap (pp)
Claude Sonnet 4.6	0.927	0.893	0.894	0.09
Claude Opus 4.7	0.903	0.890	0.892	0.20
GPT-5.4	0.834	0.783	0.786	0.34
Mistral Large	0.744	0.742	0.742	0.00
DeepSeek-R1	0.741	0.739	0.739	0.07
Gemini 2.5 Flash	0.739	0.631	0.713	8.22
Qwen3-235B	0.695	0.744	0.744	0.00
DeepSeek-V3.2	0.609	0.727	0.727	0.00

(b) Aggregation rule (cross-DB sub-tests / noisy 12-voter ensemble)

Rule	DR	FPR	F1	MCC
structured: any-miss ($k\geq$1/4)	0.950	0.000	0.975	0.948
structured: cross-DB-agreement only	1.000	0.035	0.985	0.968
structured: majority ($k\geq$3/4)	0.048	0.000	0.091	0.150
field-level: any-miss ($k\geq$1/4)	0.759	0.139	0.809	0.619
field-level: two-fields ($k\geq$2/4)	0.365	0.031	0.525	0.407
ensemble: majority ($\geq 7/12$)	0.779	0.191	0.803	0.586
ensemble: supermajority ($\geq 11/12$)	0.495	0.047	0.645	0.493

Table 46: **Input-format and field leave-one-out on `dev_public`** (n=150; temperature 0; seed 42; fresh 2026-05-31 snapshot). ΔFPR is measured against each model's structured-field baseline. Dropping the title or authors spikes FPR; dropping venue/year/DOI does not: detection leans on title/author content, not a format tell. (DOI LOO affects only the 71 DOI-bearing entries.)

Model	Condition	DR ↑	FPR ↓	F1 ↑	ΔFPR
Gemini 2.5 Flash	full	0.543	0.101	0.667	—
	structured	0.617	0.145	0.709	(base)
	−title	0.789	0.500	0.709	+0.355
	−authors	0.709	0.333	0.709	+0.188
	−venue	0.519	0.101	0.646	−0.043
	−year	0.593	0.275	0.649	+0.130
	−DOI	0.469	0.145	0.589	0.000
DeepSeek-V3.2	full	1.000	0.957	0.711	—
	structured	0.975	0.812	0.731	(base)
	−title	0.988	0.940	0.712	+0.129
	−authors	0.987	0.896	0.719	+0.084
	−venue	0.926	0.826	0.704	+0.014
	−year	0.926	0.754	0.721	−0.058
	−DOI	0.963	0.841	0.719	+0.029

human Inter-Annotator Agreement (IAA), which remains future work (§ 5); it serves to test whether independent raters reproduce the relabel.

Overall agreement is only fair: Fleiss' $\kappa = 0.24$ (Fleiss, 1971) on the "fair" band of the Landis–Koch scale (Landis & Koch, 1977), with pairwise Cohen's κ ranging from 0.45 (Sonnet–DeepSeek, moderate) down to 0.03 (DeepSeek–Gemini Pro, slight). Against the corrected ground-truth labels—the database-backed labels after the relabel audit (§ A.3)—DeepSeek-V3.2 agrees best ($\kappa = 0.716$, accuracy 0.871) and Gemini 2.5 Pro worst ($\kappa = 0.026$); the majority vote reaches accuracy 0.720 ($\kappa = 0.340$).

The diagnostic split is the point. Majority-vote accuracy is 0.975 on the 80 real-world hallucinations but only 0.327 on the 52 relabel-recovered real papers: the LLM raters confidently agree on genuine fabrications yet systematically *over-flag* the recovered real papers, the same over-flagging failure mode HALLMARK is built to measure. Independent raters would have repeated the original labeling

Table 47: **Multi-rater reliability proxy (not human IAA)** on 132 blinded `dev_public` entries (80 real-world-incident hallucinations, 52 relabel-recovered real papers), 3 independent LLM raters, binary (UNCERTAIN→HALLUCINATED). The diagnostic split does the work: the raters agree on genuine fabrications but over-flag the recovered real papers, exactly the failure mode the benchmark measures; the red-shaded cell marks that diagnostic.

Agreement statistic	**Value**	
Fleiss' κ (3 raters, binary)	0.238 (fair)	
Cohen's κ Sonnet–DeepSeek	0.454 (moderate)	
Cohen's κ Sonnet–Gemini Pro	0.266 (fair)	
Cohen's κ DeepSeek–Gemini Pro	0.029 (slight)	
Rater	**κ vs. ground truth**	**acc. vs. ground truth**
DeepSeek-V3.2	0.716	0.871
Sonnet 4.6	0.360	0.727
Gemini 2.5 Pro	0.026	0.576
Majority vote	0.340	0.720
Majority-vote accuracy by pool		
real-world hallucinations (n=80)	0.975	
relabel-recovered real papers (n=52)	0.327	

error, which both corroborates that the relabel audit corrected a real bias and motivates the precision-via-abstention contribution (§ E.2).

Takeaway. Independent raters reproduce the failure mode the benchmark measures: majority-vote accuracy is 0.975 on the 80 real-world hallucinations yet 0.327 on the 52 relabel-recovered real papers, so three independent LLM raters would have repeated the labeling error the ground-truth audit corrected. With Fleiss' $\kappa = 0.24$ overall, LLM raters are no substitute for database-backed ground truth, nor for the human IAA that remains future work (§ 5).

I Extended limitations

The seven paragraphs below are the full-length limitations discussion that § 5 ranks and compresses.

Dataset scale and coverage. The benchmark is small relative to the full diversity of citation hallucinations. At 2,526 entries, most types on `dev_public` and `test_public` clear ≥ 30 instances—enough for meaningful per-type comparison—but a few fall below it, and `test_hidden` does not reach the floor at all (§ B.1). Prevalence differs across splits (54.2–62.5%), so we emphasize the prevalence-independent metrics (DR, FPR). We cover English-language BibTeX from five ML venues (2021–2023); the taxonomy and sub-tests are designed to be domain-agnostic but validated mainly in that regime. A 500-entry cross-domain split (PubMed/bioRxiv plus non-ML CS venues), together with a recency-matched, canonically-resolved rebuild, separates domain from post-cutoff recency out of regime (§ G.2). The separation reframes the result: the released split's apparent cross-domain FPR rise is mostly a recency artifact—the LLMs flag 2026 dates as impossible—and once entries are pre-cutoff the residual domain effect is negative or small for the calibrated verifiers, large only for a single recall-aggressive model that already over-flags in its own regime (Tab. 39). For `bibtex-updater`, canonical metadata leaves the FPR at its in-domain level while coverage falls: out of domain the cost is coverage, not precision. Two failure modes stay specific to biomedical data—provider safety filters block $\sim$3% of valid biomedical citations for the Anthropic models, distinct from any verdict error—while humanities and non-English settings remain future work. Just over half the valid entries lack a DOI (53.0% on `dev_public`, 55.4% on `test_public`)—many ML papers are never formally published, and DBLP records for ICML, ICLR and NeurIPS carry no DOI at all—which raises the FPR of verifiers that treat a missing identifier as suspicious.

Synthetic–real gap. Most hallucinated entries are perturbation-generated, and the real-world anchor is thin. The benchmark holds 108 real-world entries (plus 280 LLM-generated), type-skewed (55% `plausible_fabrication`), with 15 compound-failure entries that involved judgment calls without *human* inter-annotator agreement. We report an automated three-rater reliability proxy (Fleiss' $\kappa = 0.24$, fair inter-rater agreement; § H.6) that corroborates the relabel audit but does not substitute for human IAA, which remains future work. Expanding real-world coverage from retraction databases is a priority.

Temporal fragility and the calibration question. LLM verifiers carry an implicit training-cutoff bias that we can only partly disentangle. § 4.4 measures the post-cutoff FPR rise and the later-cutoff resistance; a recall probe and a non-Anthropic late-cutoff control point away from pure memorization for Opus 4.7 (and only partly for Sonnet 4.6), but the recall measure is model-derived and residual confounds—system-prompt handling, default temperature, native-vs-OpenRouter routing—remain. We therefore read the resistance as a descriptive signal, not a causal attribution.

Controls and residual gaps. Three controls probe the calibration-versus-contamination question, none decisive (§ F.4). (i) A recall probe on valid 2024–2025 papers is consistent with calibration over memorization for Opus 4.7 (it accepts 84% of papers it cannot recall), while Sonnet 4.6's lower FPR is partly recall-driven. (ii) A third-provider late-cutoff control (DeepSeek-V4-Pro) also holds its post-cutoff FPR low relative to the over-flagging cohort (0.36 vs. 0.93 for GPT-5.1 on the same $n{=}300$ subsample, at full coverage), though its cutoff matches rather than exceeds the Anthropic pair's (§ F.4). (iii) GPT-5.1's run-to-run variance is small (F1 std 0.002 over three runs), so single-run rankings seem to be stable to sampling noise, at least for some models. The open gaps: the cross-provider control is a single third-party model, a native-API replication of the Anthropic runs is outstanding, and contamination is reduced but not fully excluded even for Opus 4.7.

Compute and token budget. Compute caps the coverage of reasoning-mode and extended-thinking variants. Reasoning models cost far more per entry (DeepSeek-R1 at a median $\sim$19 s vs. 1–10 s, Tab. 19), so we did not run extended-thinking variants on the full benchmark (Fig. 7 maps where their JSON-output contract holds), the agentic harness is capped at five tool calls per entry, and HaRC/verify-citations are excluded because neither produces its own verdicts at benchmark scale (§ B.2.2).

Regime conditionality and co-design. The "FPR decides" framing is calibrated to the reviewer-bound venue-audit regime: low prevalence, finite reviewer attention, a false alarm cheap to dismiss. The PPV *ranking* is itself prevalence-invariant—sweeping prevalence over 1–5% does not reorder the tools (Tab. 24)—so what changes across regimes is whether *any* verifier reaches a usable absolute precision and the cost asymmetry $c_{\text{FN}}/c_{\text{FP}}$; under high prevalence or a high cost of a missed fabrication, recall and human triage regain primacy. Separately, `bibtex-updater` and the taxonomy were developed in parallel, so the taxonomy reflects fundamental verification steps but the construct-overfitting risk is non-zero; we report it as a co-designed reference (§ G.2). A second co-design channel runs through the valid-entry inclusion rule, and it is arguably stronger than the taxonomy: § 3.2 admits a valid entry only after confirming DOI resolution, title existence in at least two databases, and author–venue consistency, so every valid entry is database-confirmable by construction, and FPR—the axis we rank on—is computed on exactly that half. The rule therefore favors database-lookup architectures structurally, independently of how the taxonomy was derived. What bounds the effect is that it does not decide the outcome: an independent parametric verifier matches it: Gemini 2.5 Pro's FPR ($.050$) is level with the co-designed tool's ($.051$). Whether rankings on synthetic hallucinations predict performance on real errors remains open (Hardt, 2026); the 108 real-world entries are an initial signal, and the codebase tracks the number of times the dev split has been evaluated against during development, an adaptive-data-analysis hygiene measure that bounds the overfitting introduced by repeatedly reusing the same held-out set (Dwork et al., 2015).

Source availability is a second measurement condition. Every number a retrieval-dependent verifier produces is conditional on whether the databases answered, and we report none of that conditioning per row. It is not a small effect. Holding tool and split fixed and correcting only the source conditions—keys, credentials and a 20-per-minute pace—moved the fraction of entries with at least one lookup that never completed from 53.7% to 10.0%, and the run got *faster* rather than slower, 2.4 s per entry against 5.7: a starved source turns into a retry storm that costs more than the wait it avoids. On a 119-entry probe holding model and prompt fixed, 12% of flags cleared with arXiv starved against 24% with it answering: the same verifier on the same entries, half the corrections. Nothing in this paper controls for it: our rows were collected under whatever the sources were doing that day, which bounds how exactly any of them can be reproduced, and we flag it as an evaluation condition that citation-verification benchmarks, ours included, do not yet report.

Endpoint drift and reproducibility. Several ablations and the two Anthropic risk–coverage curves were collected against the OpenRouter API on a dated snapshot that does not reproduce the main-run aggregates. The Anthropic endpoints drift over time, and for these two models the drift is large enough to dominate several of the effects we measure. Two independently collected later runs agree on its direction and rough size: an OpenRouter snapshot puts Opus 4.7 at $0.072{\to}0.162$ and Sonnet 4.6 at

$0.127 \rightarrow 0.165$, and the native 2026-09-02 re-run lands at 0.154 and 0.161 (Tab. 29). Both roughly double the May FPR under a fixed model ID, prompt, split, temperature and seed. That re-run also persists per-entry predictions, which the original main run did not, so the Sonnet/Opus coverage cells and dev confidence intervals are now recoverable and are reported rather than marked "n/a". Our convention follows from the size of the drift rather than from a preference for either collection. Cross-sectional comparisons take the latest run, so Tab. 1 reports 2026-09-02 for these two rows. Any comparison whose quantity is a *difference* stays inside one collection, because a 13 pp endpoint shift swamps the differences at stake: the post-cutoff shift in Tab. 28 is under 1 pp for both models, so those rows remain wholly on May, and the agentic-inflation ratio of failure mode (i) is likewise quoted within May. Mixing the two would not merely add noise, it would report drift as a finding—pairing a September 21–23 FPR with a May 24–25 one yields -4.1 and -8.1 pp, i.e. apparent *improvement* on post-cutoff papers. Rankings are the more robust reading in any case, since drift shifts absolute FPR in the same direction across models while largely preserving their order (cross-regime ranking at Spearman ρ=0.90 under prompt perturbation, § H.2). Within-run deltas (prompt-variant, field-LOO, rater-agreement) difference conditions measured against the same endpoint on the same day and so survive the drift. The released corpus is pinned at tag `v1.2.0`, carrying the ground-truth audit of § A.3.[7]

Diagnostic resolution is lower than six. Three of the six sub-tests carry little discriminative load on this corpus. Cross-database agreement is expected-fail for all 14 hallucination types and expected-pass for every valid entry, so it separates the two classes perfectly and localizes nothing; `fields_complete` is expected-fail for no type and varies only for `plausible_fabrication`; and `doi_resolves` is not applicable, or varies by entry, for seven of the fourteen (Tab. 2). The effective number of sub-tests that discriminate *between* hallucination types is therefore nearer three than six, which bounds how finely a failure signature can attribute an error and is a target for the next taxonomy revision.

J BROADER IMPACT AND DEPLOYMENT SAFEGUARDS

HALLMARK supports scientific integrity by enabling systematic comparison of citation-verification tools, informing venue deployment as LLM-assisted writing becomes standard. A taxonomy of hallucination types could theoretically aid adversaries, but the defensive value of understanding these types outweighs the risk; the taxonomy is derived from publicly documented incidents, not novel attack research. The benchmark is open-source (MIT) and GPU-free, running on a single CPU regardless of institutional resources.

Intended use, and two uses we advise against. HALLMARK is a diagnostic for tool builders and venue operators: it measures which verifier fails on which citation error, and at what precision cost. It is not built for autonomous rejection of submissions, nor for screening authors for research misconduct, and we advise against both. Both inferences require establishing intent, which no verifier in this evaluation attempts and no metadata check can support: a flag means the metadata did not resolve against the databases queried, not that an author fabricated anything. Our own measurements are the argument against automating the decision. At a venue-realistic $\sim 2\%$ base rate the best verifiers in the cohort catch one true hallucination per 6–9 flags and the most aggressive fewer than one in 35 (Tab. 23), so no operating point we measured would make an automated gate defensible.

Deployment safeguards. Three safeguards follow from those measurements rather than from principle. A human decides: a verifier's output routes an entry to a reviewer for checking and never stands as a verdict on the citation. A flag arrives with its evidence—which databases were queried and what each returned—so that an author can rebut it without reverse-engineering the tool. And an appeal route exists and is cheap: an author who supplies a resolving identifier or a canonical database record clears the flag without adjudication, which is the correct outcome for the recency and indexing failures below, where the tool is wrong and the citation is fine.

[7] Later data releases (`v1.2.1`, `v1.2.2`) replace stale generation-pipeline explanation strings with diagnoses verified against CrossRef/arXiv/DBLP records (per-entry fix log in the repository); labels, hallucination types, tiers, sub-tests, and split membership are byte-identical to `v1.2.0`. Tag `v1.2.3` additionally retypes four entries to `plausible_fabrication`/Tier 3 (`ffe715a15b16` from `fabricated_doi` and `ec01d96455e0` from `preprint_as_published` on `dev_public`; `ab9c13051a56` and `cad998c28243` from `placeholder_authors` on `test_public`) and moves the baseline results to `data/v1.2/`. Split membership and VALID/HALLUCINATED labels are unchanged, so aggregate DR, FPR and F1 are unaffected; the per-type and per-tier counts in this paper are those of `v1.2.3`.

Who a false positive lands on. False alarms are not distributed evenly across authors, and the unevenness is measurable rather than speculative. Averaged over the eighteen verifiers with released per-entry predictions, FPR orders the five venues identically on both public splits—ICML 0.44, ICLR 0.36, NeurIPS 0.32, AAAI 0.13, CVPR 0.09—and the mean pairwise Spearman correlation between verifiers' per-venue rates is $+0.56$, so the burden is a property of the corpus rather than of any one tool. What the ordering tracks is bibliographic completeness, not venue standing: no valid ICML, ICLR or NeurIPS entry in the public splits carries a DOI, against every entry at CVPR and AAAI. Per-verifier spread across those venues runs from 1.3 pp for Gemini 2.5 Pro and 6.3 pp for `bibtex-updater` to 69.5 pp for GPT-5.1 and 93.9 pp for the agentic harness, so a low-FPR verifier also buys uniformity, while an aggressive operating point concentrates its errors on whoever is least well indexed. Publication recency behaves as a cliff rather than a gradient: within 2021–2023 no verifier's year-to-year spread exceeds 19 pp, but on the 35 real February-2026 arXiv preprints in the corpus—post-cutoff for every model evaluated—all eighteen verifiers over-flag and fifteen reach FPR 1.000, and on the 448-entry 2024–2025 supplement all twelve LLMs are worse on 2025 than on 2024 (§§ 4.4 and F.1). The authors most exposed are therefore those publishing recent work and work whose venue records carry no resolvable identifier, which is close to the inverse of who can most easily absorb the cost. A venue that deploys a verifier without first measuring its own subgroup false-positive rates will impose this burden without observing it.

Scope of the evidence. The conclusions here are established on English-language BibTeX from five ML venues (2021–2023), extended by a 500-entry cross-domain split covering biomedical and non-ML computer-science sources and by an authentic-LLM-output supplement spanning 42 multidisciplinary topics (§§ C.6 and G.2). They are not established for non-English citation practices, for humanities conventions where monographs and archival sources dominate, or for fields whose canonical identifiers differ from the DOI/arXiv pair these tools query. Deploying a verifier in those settings requires re-measuring its false-positive rate there first, and this benchmark does not license the transfer.

K ABBREVIATIONS

AUROC Area Under the ROC Curve

DOI Digital Object Identifier
DR detection rate

ECE Expected Calibration Error

F1 F1 score
FPR false positive rate

GEPA Genetic-Pareto

IAA Inter-Annotator Agreement

LLM Large Language Model
LOO leave-one-out

MCC Matthews Correlation Coefficient
MDE Minimum Detectable Effect

PF parse-failure rate
PPV positive predictive value

RLHF Reinforcement Learning from Human Feedback

TW-F1 tier-weighted F1