Re^{2}Math: Benchmarking Theorem Retrieval in Research-Level Mathematics

Zicheng Lyu [*1] **Wenjie Yang** [*1] **Shengzhong Zhang** [2] **Zengfeng Huang** [1 3]

Abstract

Large language models are increasingly capable at closed-world mathematical reasoning, but research assistance also requires source-grounded use of the literature. When a proof reaches a non-trivial step, a useful assistant should determine whether the needed tool (e.g., a lemma) already exists, identify a suitable scholarly source, and verify that its assumptions align with the current proof context. To rigorously evaluate such capabilities, we introduce Re^{2}Math, a benchmark for *tool-grounded retrieval* from partial mathematical proofs. Each instance is built from a candidate instrumental citation in the proof of a main theorem, with hierarchical context and an optional leakage-controlled anchor hint. We also make the task source-grounded yet citation-agnostic in that any admissible theorem sufficient for the proof transition is accepted. Evaluation uses a release-frozen retrieval artifact, ensuring reproducibility, while the benchmark itself supports automatic, continual expansion with newly constructed instances. On the current benchmark test set, the best fixed-judge ToolAcc reaches 7.0%, despite substantially higher rates of source grounding, indicating that current systems often retrieve valid statements but fail to establish their applicability to the local proof step. By decoupling citation recall, grounding, and proof-gap sufficiency, Re^{2}Math transforms literature-grounded mathematical tool use into a controlled diagnostic task.

1. Introduction

Large language models (LLMs) have improved rapidly on mathematical reasoning benchmarks, from grade-school arithmetic to olympiad-style problem solving (Cobbe et al., 2021; Hendrycks et al., 2021; He et al., 2024; Gao et al., 2024a; Sun et al., 2025; Glazer et al., 2024). Yet a useful mathematical research assistant must do more than solve self-contained problems. Research proofs often reach steps that require an external theorem, estimate, compactness criterion, representation result, or gluing principle. At such points, the assistant should locate prior work, identify an admissible source, and check whether the source's assumptions apply to the current proof state.

This motivates *tool-grounded retrieval*. A model may state a theorem-like claim from memory, but a claim without provenance leaves open where the result was established, how it should be attributed, and whether its hypotheses match the proof at hand. The object to be evaluated is therefore a source–tool pair: a source that supports the proposed mathematical tool, and a statement whose assumptions and conclusion are sufficient for the local proof transition. This is different from both unsupported theorem suggestion and exact citation recovery.

Existing benchmarks only partially capture this capability. Outcome-oriented math datasets ask for a final answer or complete solution from a largely self-contained prompt, while formal proving benchmarks evaluate search inside fixed formal libraries (Zheng et al., 2021; Azerbayev et al., 2023; Tsoukalas et al., 2024). They do not isolate open-world literature use from partial research-proof context, nor do they test whether a model can turn a proof gap into a verifiable source–tool pair.

We introduce Re^{2}Math, a benchmark for tool-grounded retrieval. We derive proof gaps from citations that our construction pipeline identifies as instrumental: citations that appear to supply an external mathematical tool for a local transition inside a main-theorem proof. Each instance supplies hierarchical context—a global setup and a local proof window immediately preceding the citation—and an assisted track additionally provides a sanitized anchor hint. Given this context, an agent must infer the proof need, generate search queries, select a source, and extract or faithfully restate a theorem-like statement. Appendix B audits this construction signal and reports the remaining risks in instrumentality and anchor sanitization.

The primary leaderboard uses controlled, model-authored

[*]Equal contribution [1]Fudan University, Shanghai, China [2]Nanjing University of Aeronautics and Astronautics, Nanjing, Jiangsu, China [3]Shanghai Innovation Institute, Shanghai, China. Correspondence to: Zengfeng Huang <huangzf@fudan.edu.cn>, Zicheng Lyu <lyuzicheng@gmail.com>.

The 3rd *AI for Math Workshop* at the 43rd International Conference on Machine Learning (ICML), Seoul, South Korea, 2026. Copyright 2026 by the author(s).

retrieval. Models may propose their own search queries, but query execution is performed through a shared release-frozen artifact rather than through model-native web-search, browser, file-reading, or deep-research modes. This restriction does not remove retrieval from the task; it fixes the information-access boundary so that hidden ranking, browsing, source-opening, and context-packaging differences do not dominate the mathematical source–tool policy being evaluated. Re^{2}Math therefore asks whether a system can produce a verifiable scholarly justification under a frozen release, not whether it can reproduce the original author's historical search process. Appendix E details this benchmark contract.

A central design choice is that Re^{2}Math is source-grounded but citation-invariant. The author-cited paper is useful for construction and diagnostics, but it is not the definition of correctness. A system receives credit whenever it returns any admissible source-grounded statement that closes the gap. This matches research practice, where the same tool may appear in multiple papers, books, surveys, or stable preprints, and a stronger alternative theorem may suffice if its assumptions are satisfied.

Re^{2}Math is built by an automated published-first, arXiv-backed pipeline and released as a versioned snapshot of a living benchmark. We evaluate source-grounded mathematical tool retrieval under a fixed-budget, shared-backend scholarly-search setting. In this setting, the best evaluated model on the benchmark test set (Eval-200) reaches only 7.0% ToolAcc, despite substantially higher grounding rates. We treat the leaderboard as a diagnostic rather than a ranking endpoint: the main empirical signal is that current systems often find or restate real mathematical claims without making them usable for the exact proof transition.

Our contributions are summarized as follows:

- **Task formulation.** We formalize tool-grounded retrieval from partial research proofs with hierarchical context, raw and assisted tracks, and a citation-invariant success criterion.
- **Living benchmark.** We construct a versioned benchmark from instrumental proof citations via a published-first, arXiv-backed pipeline that *automatically* incorporates new instances as the literature grows, while keeping paper-facing evaluation fixed to a frozen held-out release.
- **Diagnostic evaluation.** We provide a shared-backend protocol with *ToolAcc* as the primary metric and diagnostics that separate citation-side retrieval, source grounding, proof-gap sufficiency, oracle-source behavior, and alternative-source success.

2. Related Work

Mathematical reasoning and proof benchmarks. GSM8K and MATH established standard benchmarks for multi-step mathematical reasoning (Cobbe et al., 2021; Hendrycks et al., 2021); later datasets pushed toward harder, refreshed, or leakage-resistant evaluation, including OlympiadBench, Omni-MATH, OlymMATH, FrontierMath, HARP, LiveBench, and LiveAoPSBench (He et al., 2024; Gao et al., 2024a; Sun et al., 2025; Glazer et al., 2024; Yue et al., 2024; White et al., 2024; Mahdavi et al., 2025). Formal benchmarks such as miniF2F, ProofNet, and PutnamBench evaluate proof production in mechanically checkable settings (Zheng et al., 2021; Azerbayev et al., 2023; Tsoukalas et al., 2024). These benchmarks measure closed-world problem solving, proof search, or formalization. Re^{2}Math targets a different research-assistance primitive: from a partial research proof, a system must identify an external mathematical tool, ground it in an admissible scholarly source, and check that it suffices for the local proof transition.

Inference-time reasoning and controlled retrieval. Modern reasoning systems increasingly rely on inference-time computation and verification, from Chain-of-Thought, self-consistency, and Tree of Thoughts to frontier and verifier-guided mathematical systems (Wei et al., 2022; Wang et al., 2022; Yao et al., 2023; OpenAI, 2024; Guo et al., 2025; Wang et al., 2024; Xin et al., 2024; Hubert et al., 2025). Retrieval and attribution benchmarks such as KILT, BEIR, and ALCE study shared retrieval settings, external-memory access, and citation-supported generation (Petroni et al., 2021; Thakur et al., 2021; Gao et al., 2023). Re^{2}Math is complementary: rather than proposing a new reasoning algorithm or a general RAG benchmark, it fixes a controlled retrieval interface and evaluates the mathematical policy mapping a proof state to a grounded source–tool pair.

Scientific literature use and citation recommendation. Scientific claim-verification benchmarks such as SciFact and SciFact-Open retrieve evidence from papers to support or refute scientific claims (Wadden et al., 2020; 2022). Literature assistants such as PaperQA and OpenScholar search scientific papers to produce citation-backed answers (Lála et al., 2023; Asai et al., 2024), while SPECTER and SciRepEval evaluate scientific document representations for search, ranking, and recommendation (Cohan et al., 2020; Singh et al., 2023). Context-aware citation recommendation retrieves papers from a local writing context (He et al., 2010; Färber & Jatowt, 2020). These works connect textual contexts to scholarly sources, but typically target article recommendation, claim evidence, or literature synthesis. In contrast, Re^{2}Math uses a mathematical proof state as context and requires a theorem-like tool whose assumptions and

conclusion are sufficient for a specific proof gap, not merely topically relevant.

Theorem retrieval and premise selection. Retrieval is also central to mathematical reasoning. Natural-language resources such as NaturalProofs and TheoremQA support theorem-conditioned reasoning and theorem use (Welleck et al., 2021; Chen et al., 2023); formal premise-selection systems and benchmarks such as MPTP, DeepMath, HolStep, HOList, LeanDojo, LeanSearch, and Magnushammer retrieve accessible premises from a formal library or proof state (Urban, 2006; Alemi et al., 2016; Kaliszyk et al., 2017; Bansal et al., 2019; Yang et al., 2023; Gao et al., 2024b; Mikuła et al., 2023). Re^{2}Math differs along the evaluation axis: retrieval is mediated by a frozen scholarly-search interface over an open-world scholarly source space; the output is a sourced theorem-like statement rather than a premise identifier or relevance score; and success is citation-invariant—any admissible source receives credit if it grounds a statement sufficient for the proof gap.

3. Task Definition

We study *tool-grounded retrieval*. Given the proof state immediately before a non-trivial proof transition and access to a scholarly search interface over admissible sources, an agent must return a source–tool pair

$$(\hat{d}_i, \hat{K}_i),$$

where $\hat{d}_i$ is an admissible source and $\hat{K}_i$ is a theorem-like statement grounded in that source and sufficient for the transition. The task is *source-grounded* but *citation-invariant*: the selected source must support the returned statement, but it need not be the exact source cited by the original author. The author citation is one human witness for the gap, not the definition of correctness.

This formulation separates Re^{2}Math from both unsupported theorem generation and citation classification. A successful system must infer the proof need, express it in searchable language, select a source from noisy candidates, and extract a statement whose assumptions fit the local context. We expose these stages for diagnosis while scoring a single end-to-end object: the returned source–tool pair. A high-level overview of the task is illustrated in Figure 2.

3.1. Instances and Input Tracks

Let a proof be segmented into ordered blocks

$$P = (s_1, \dots, s_T),$$

where each block is a sentence, displayed equation, or mathematical environment. Suppose a citation occurrence in block s_c is treated as a witness for a non-trivial proof transition. We expose only the pre-citation state $k = c - 1$, and write Δ_i for the target transition into the citation block.

Each instance i provides hierarchical context

$$\mathcal{C}_i = (\mathcal{C}^i_{\text{global}}, \mathcal{C}^i_{\text{local}}(m)),$$
$$\mathcal{C}^i_{\text{local}}(m) = (s_{\max\{1,k-m+1\}}, \dots, s_k).$$

The global context summarizes the paper setup, assumptions, definitions, and target theorem; the local context is the ordered proof window immediately before the citation. The canonical stored window keeps up to five pre-citation blocks, and local-window diagnostics use $m \in \{1, 3, 5\}$.

Each instance may also include a sanitized anchor a_i, a short description of the immediate proof role of the missing tool. The anchor is a planning cue, not a bibliographic shortcut: it should not contain theorem titles, author names, citation keys, DOIs, arXiv identifiers, or source metadata. We define two input tracks:

$$\mathsf{x}^i_{\text{raw}} := (\mathcal{C}^i_{\text{global}}, \mathcal{C}^i_{\text{local}}),$$
$$\mathsf{x}^i_{\text{assist}} := (\mathcal{C}^i_{\text{global}}, \mathcal{C}^i_{\text{local}}, a_i).$$

The Raw track requires the agent to infer the proof need from context alone. The Assisted track fixes this planning cue so that retrieval, source selection, and extraction can be studied more directly.

For supervision and analysis, each instance stores

$$\underbrace{\mathsf{x}_i}_{\text{agent input}} \qquad \underbrace{\mathsf{y}_i = K^\star_i}_{\text{reference tool witness}} \qquad \underbrace{\mathsf{z}_i = d^{\text{cite}}_i}_{\text{citation witness}}.$$

Here $K^\star_i$ is one curated theorem-like witness aligned to the gap, and d^{cite}_i is the source used by the original author. Together they certify that the transition is externally tool-dependent, but they do not exhaust the valid answer set.

3.2. Retrieval Interface and Agent Policy

Let $\mathcal{D}$ be the admissible scholarly source space and let $\mathcal{K}(d)$ denote theorem-like statements stated in, or faithfully entailed by, a source $d \in \mathcal{D}$. We model retrieval abstractly by an operator

$$\rho_N : \mathcal{Q}^L \to \mathcal{D}^N,$$

which maps a bounded query list to a ranked candidate prefix of length at most N. The experimental protocol instantiates ρ_N with a shared release-frozen retrieval interface; the task definition only assumes that all agents use the same retrieval operator.

An agent proceeds through four conceptual stages:

$$\mathsf{x}_i \xrightarrow{\pi_{\text{plan}}} \hat{a}_i \xrightarrow{\pi_{\text{qry}}} Q_i \xrightarrow{\rho_N} \mathcal{R}_{i,N} \xrightarrow{\pi_{\text{sel}}} \hat{d}_i \xrightarrow{\pi_{\text{ext}}} \hat{K}_i.$$

This diagram is a workflow factorization, not a strict unary composition: selection and extraction may condition on the proof context $\mathcal{C}_i$ and the planned proof need $\hat{a}_i$. More explicitly,

$$\hat{a}_i = \begin{cases} \pi_{\text{plan}}(\mathsf{x}^i_{\text{raw}}), & \text{Raw track,} \\ a_i, & \text{Assisted track,} \end{cases}$$
$$Q_i = \pi_{\text{qry}}(\mathcal{C}_i, \hat{a}_i),$$
$$\mathcal{R}_{i,N} = \rho_N(Q_i),$$
$$\hat{d}_i = \pi_{\text{sel}}(\mathcal{C}_i, \hat{a}_i, \mathcal{R}_{i,N}) \in \mathcal{R}_{i,N} \cup \{\perp\},$$
$$\hat{K}_i = \pi_{\text{ext}}(\mathcal{C}_i, \hat{a}_i, \hat{d}_i) \in \mathcal{K}(\hat{d}_i) \cup \{\perp\}.$$

The stages diagnose distinct capabilities: recognizing the missing mathematical role, lexicalizing it for search, choosing a source, and extracting a usable result. Only the final pair

$$\hat{\mathsf{y}}_i = (\hat{d}_i, \hat{K}_i)$$

is the official task output.

3.3. Valid Outputs

A source–tool pair is valid if the statement is grounded in an admissible source and sufficient for the transition Δ_i. Admissible sources are scholarly documents external to the local citing proof: the citing paper itself, and its arXiv/published variants, do not count merely by restating the citation context or the proof sentence that generated the instance. They are admissible only if the selected text independently states or proves the external tool being invoked.

We define the valid-output set

$$\mathcal{V}_i := \{(d, K) \in \mathcal{D} \times \mathcal{K}(d) : \mathsf{Ground}_i(d, K) = 1 \\ \wedge\ \mathsf{Suff}_i(K \mid \mathcal{C}_i, \Delta_i) = 1\}.$$

$\mathsf{Ground}_i(d, K)$ checks whether K is supported by d, either verbatim or as a mathematically faithful restatement. Grounding is a provenance predicate, not an independent verification that the source theorem is true. $\mathsf{Suff}_i(K \mid \mathcal{C}_i, \Delta_i)$ checks whether K, together with the proof context, justifies the transition; stronger statements receive credit only when their additional assumptions are satisfied. Thus correctness is membership in $\mathcal{V}_i$, not equality to d_i^{cite}.

3.4. Metrics

For an evaluation split $\mathcal{T}$, the primary metric is

$$\text{ToolAcc}(\pi; \mathcal{T}) := \frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{T}} \mathbf{1}[(\hat{d}_i, \hat{K}_i) \in \mathcal{V}_i].$$

We also report a Raw-track planning diagnostic, AnchorAcc. Let $\widehat{a}_i^{\text{raw}}(m) := \pi_{\text{plan}}\big(\mathcal{C}^i_{\text{global}}, \mathcal{C}^i_{\text{local}}(m)\big)$ be the proof-intention description produced before retrieval, and let $\mathsf{AnchorMatch}_i(\widehat{a}, a_i) = 1$ mean that $\widehat{a}$ captures the same immediate proof role as the stored sanitized anchor a_i, up to benign paraphrase. Then

$$\text{AnchorAcc}_m(\pi; \mathcal{T}) := \frac{1}{|\mathcal{T}_a|} \sum_{i \in \mathcal{T}_a} \mathbf{1}\left[\begin{matrix}\mathsf{AnchorMatch}_i(\\ \widehat{a}_i^{\text{raw}}(m), a_i) = 1\end{matrix}\right],$$
$$\mathcal{T}_a := \{i \in \mathcal{T} : a_i \neq \perp\}.$$

The main paper reports $\text{AnchorAcc} := \text{AnchorAcc}_5$ with global context and the five-block local window. It measures proof-need abstraction before retrieval and is not a success criterion.

We report the grounding rate

$$\text{GroundRate}(\pi; \mathcal{T}) := \frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{T}} \mathbf{1}[\mathsf{Ground}_i(\hat{d}_i, \hat{K}_i) = 1],$$

which measures source-supported extraction without requiring proof-gap sufficiency.

Finally, to analyze citation-side retrieval without making it the target, let $d \equiv_{\text{src}} \mathsf{z}_i$ denote source identity between a retrieved candidate and the stored citation metadata. Then

$$\text{CiteRecall@}N := \frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{T}} \mathbf{1}[\exists d \in \mathcal{R}_{i,N} : d \equiv_{\text{src}} \mathsf{z}_i].$$

CiteRecall@N asks whether the author-used construction witness appears in the retrieved set. It remains diagnostic only: the official correctness criterion is $(\hat{d}_i, \hat{K}_i) \in \mathcal{V}_i$. Implementation details for the retrieval contract, source-identity relation, judging rubric, and auxiliary diagnostics are given in Section 5 and Appendix C.

4. Benchmark Construction

Re^{2}Math instantiates the task in Section 3: each row stores a proof state $\mathcal{C}_i$, an optional anchor a_i, a reference witness $K_i^\star$, and citation metadata d_i^{cite}; evaluation asks for any valid source–tool pair $(\hat{d}_i, \hat{K}_i) \in \mathcal{V}_i$. We summarize the construction, snapshot, and domain coverage here, and defer prompts, matching logic, curation rules, release details and anonymous release URL to Appendix B.

4.1. High-Level Pipeline

The pipeline is *published-first, arXiv-backed*: we select proof-rich published papers, preserving their scholarly identity and citation structure, and use high-confidence arXiv twins with LaTeX source when available to localize theorem statements, proof spans, and citation occurrences. This keeps the benchmark grounded in published mathematical writing while making extraction repeatable and refreshable.

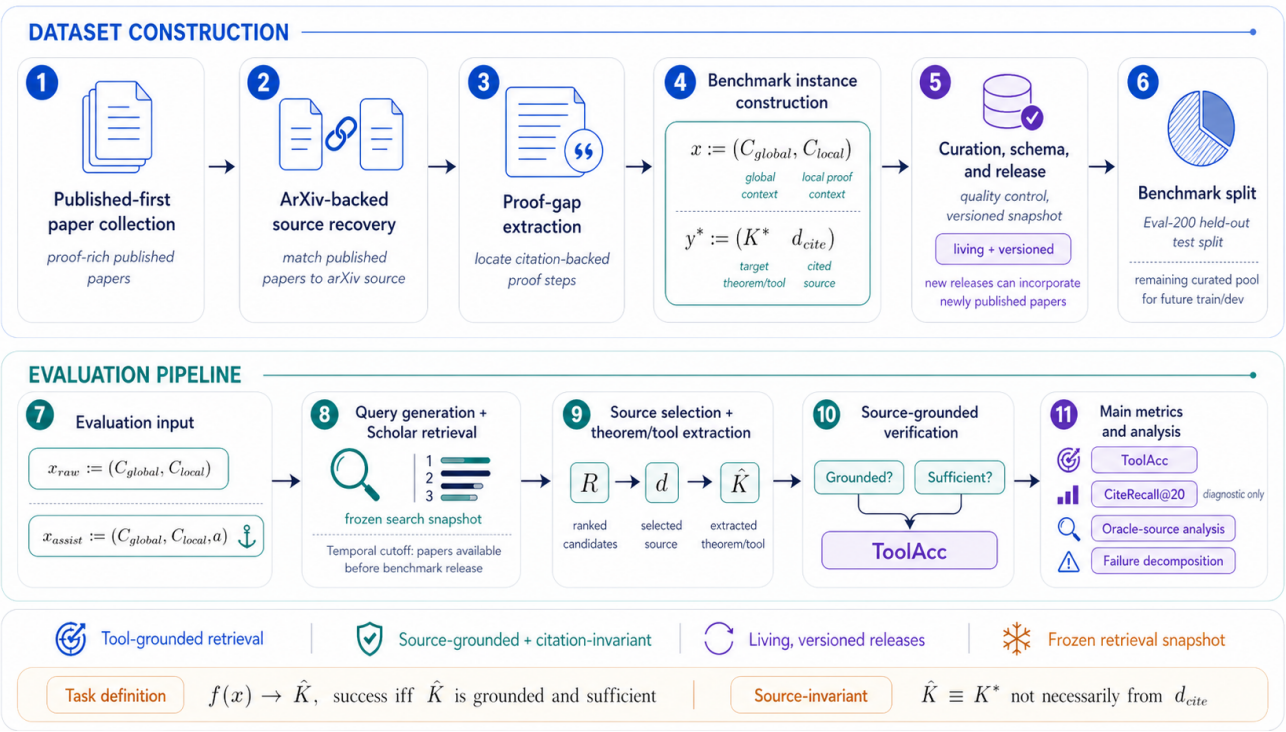

Figure 1. **Overview of Re^{2}Math.** A published-first, arXiv-backed pipeline mines candidate instrumental citations inside main-theorem proofs and converts them into proof-gap instances. Evaluation follows planning or anchor use, query generation, source selection, and theorem/tool extraction against a release-frozen retrieval artifact.

For each accepted published–arXiv pair, a paper-level stage identifies the setup, target theorem, main-theorem proof span, and candidate instrumental citations. A citation-level stage converts retained citation occurrences into benchmark instances by extracting the pre-citation proof state $\mathcal{C}_i$, a leakage-controlled anchor a_i, a reference witness $K_i^\star$, and auxiliary citation metadata d_i^{cite}. Ambiguous matches, low-confidence alignments, and suspicious witnesses are filtered during curation rather than forced into the release.

4.2. Instance Format and Running Example

Each instance stores the formal objects used by the task: the agent input x_i, the reference witness $K_i^\star$, and auxiliary citation metadata d_i^{cite}. The input consists of global context $\mathcal{C}^i_{\text{global}}$, a canonical local proof window $\mathcal{C}^i_{\text{local}}(5)$, and optionally the anchor a_i for the Assisted track. Retrieval-time artifacts such as Q_i, $\mathcal{R}_{i,N}$, $\hat{d}_i$, and $\hat{K}_i$ are produced by the evaluated agent, not stored as labels.

Running example. In one algebraic K-theory instance, `2110.11087v2_gap_1`, the global context defines $K_2^G(R) = \ker(\mathrm{St}(\Phi, R) \to G(R))$ for Chevalley–Demazure groups and aims to prove $K_2^G(R[t]) \cong K_2^G(R)$. The local proof has handled axioms CFC and DP and leaves a type-A gluing step. The Assisted anchor asks for the polynomial gluing property of Steinberg groups; the reference witness $K_i^\star$ states that $\mathrm{St}(A_\ell, R[t])$ satisfies this property for regular $R \supset k$, $\ell \geq 4$; and the author-cited witness is Tulenbaev's 1983 paper. A prediction succeeds only if it returns an admissible source supporting a sufficient gluing statement, not necessarily the citation key `Tu83`. Appendix F.9 gives the full normalized record.

4.3. Current Snapshot and Domain Coverage

The current release provides a snapshot of Re^{2}Math, containing 860 proof-gap instances from 631 distinct papers after arXiv-version normalization. The dataset spans five domains: geometry/topology, analysis/PDE, probability/statistics/control, algebra/number theory, and combinatorics/discrete mathematics. For evaluation, we construct a balanced Eval-200 test split with 40 instances per domain. Eval-200 is quality-controlled rather than random: it prioritizes journal-paper witnesses, nonempty local context, anchors subjected to leakage controls, and complete reference witnesses; Appendix B.6 reports a separate construction audit of residual citation-role and anchor-sanitization risks. Re^{2}Math is designed as a versioned, continuously expanding benchmark. Future releases may extend the instance set and domain coverage by incorporating newly published papers through the same automated pipeline. *The pipeline is*

designed to be largely automated and low-friction, supporting incremental expansion with minimal additional curation effort. An overview of the current dataset statistics is illustrated in Figure 3.

5. Experiments

The experiments instantiate the source–tool contract in Section 3. We evaluate a fixed-budget research-assistance diagnostic: can a model map a proof state to a verifiable mathematical tool when all systems share the same information-access boundary? Models write the queries, but all retrieval is executed through the same release-frozen backend. This keeps retrieval inside the task while preventing model-native web-search, browser, file-search, or deep-research stacks from becoming hidden confounders.

5.1. Evaluation Conditions

Canonical end-to-end. The primary condition is the Raw track with global context and the stored five-block local window. For each instance, the model infers the proof need, emits search queries, observes the Top-20 metadata returned by the shared backend, selects at most one resolvable source to open in full text, and returns $(\hat{d}_i, \hat{K}_i)$. This is the main leaderboard condition because it tests the complete policy: planning, query generation, source selection, and theorem/tool extraction. In this condition, a completed run uses one Scholar query and at most one full-text source opening per instance; uncertain, abstained, or non-evaluable judge outcomes are counted as negative.

Oracle-source diagnostic. To separate source acquisition from source use, we also run an oracle-source condition. When the author-cited source can be resolved and materialized under the release protocol, the model receives that source directly and must still localize, restate, and apply the relevant tool. $\mathrm{OracleCoverage}$ is the fraction of Eval-200 for which this cited source is materializable, and oracle $\mathrm{ToolAcc}$ is computed only on that subset. This diagnostic probes within-source reading and sufficiency checking; it is not a strict upper bound, because end-to-end systems may legitimately succeed through different admissible sources.

Auxiliary diagnostics. Appendix D reports planning-window and assisted-query ablations, uncertainty intervals, domain stratification, source-level failure decomposition, and a single-annotator audit of judge labels. These diagnostics explain where failures enter the pipeline, but they do not replace the primary metric: whether the final source–tool pair belongs to $\mathcal{V}_i$.

5.2. Common Setting

We evaluate on the official held-out Eval-200 split of the current release. The remaining curated instances are reserved as a non-test pool for future train/dev use and are not used for training, prompt tuning, or model selection in this paper. Splits are defined at the citing-paper level, so multiple gaps from the same paper cannot cross partitions.

Models. The paper-facing results report seven completed model runs: GPT-5.2, Gemini 3.1 Pro, Claude Opus 4.5, DeepSeek V3.2, Qwen3-235B Thinking, Kimi K2 Thinking, and Grok 4. Each reported run uses the same benchmark interface, retrieval budget, source-materialization policy, and judge rubric.

Shared retrieval artifact. All models use the same release-frozen retrieval artifact, instantiated with Google Scholar. We use Google Scholar as a broad scholarly discovery interface because the needed tool may appear across heterogeneous records—articles, proceedings papers, preprints, books, surveys, or bibliographic entries—rather than in an arXiv-only or publisher-specific index. For each proof gap, model-authored queries are executed once under fixed externally controlled Google Scholar settings, with no lower-bound year and an upper-bound year given by the original citing paper's publication year. The evaluator caches the returned Top-20 metadata, rank order, query timestamp, and source-materialization outcomes when attempted. The benchmark therefore fixes two things: a per-instance bibliographic cutoff to avoid later-literature leakage, and a frozen query-result artifact for reproducibility. Since Google Scholar's underlying coverage and ranking are inherited rather than re-indexed by us, $\mathrm{CiteRecall@20}$ is a query–artifact diagnostic, while $\mathrm{ToolAcc}$ remains the primary source–tool metric.

Judge. Outputs are scored by a held-out non-roster judge model fixed across runs. The judge applies the two predicates in Section 3: grounding in the selected source and sufficiency for the proof transition. Appendix C gives the prompt rubrics, and Appendix D reports uncertainty diagnostics together with a single-annotator human audit. The audit is judge-vs-human-confirmed validation, not an inter-annotator agreement study.

Metric. The main paper reports the primary end-to-end metric, $\mathrm{ToolAcc}$, together with three diagnostics: $\mathrm{AnchorAcc}$, $\mathrm{CiteRecall@20}$, and $\mathrm{GroundRate}$ defined in Section 3.4.

The oracle table additionally reports $\mathrm{OracleCoverage}$, $\mathrm{Oracle\ ToolAcc}$, the oracle gap, and alternative-source diagnostics. $\mathrm{OracleCoverage}$ is the fraction of Eval-200 instances for which the author-cited source could be re-

solved and materialized in a completed oracle run; it is a release-protocol materialization property, not a model score. Oracle ToolAcc is computed on the materialized subset for each completed run and should be read as an upper-bound-style diagnostic rather than a perfectly controlled same-subset counterfactual. We read the metrics as a staged diagnosis: AnchorAcc probes proof-need abstraction, CiteRecall@20 probes whether the stored witness appears in the retrieved pool, GroundRate probes source-supported extraction, ToolAcc probes mathematical sufficiency, oracle evaluation probes within-source reading, and alternative-source metrics test the citation-invariant design.

6. Results and Analysis

Under the fixed-budget shared-backend setting, current frontier LLMs are not reliable at tool-grounded mathematical retrieval. The strongest fixed-judge score is only 7.0% ToolAcc, and all other fixed-judge scores fall between 1.0% and 3.5%. We therefore treat the leaderboard as a diagnostic rather than a ranking endpoint: the main signal is the failure structure. Table 1 already shows this structure. Kimi has the strongest planning signal, Grok has the highest grounding rate, but Claude has the best end-to-end ToolAcc. Intermediate progress is not a monotone proxy for source–tool success.

Grounding is much easier than applicability. Models often retrieve or restate true mathematics without making it usable. Grok grounds statements on $83/200$ instances but succeeds on only $7/200$; DeepSeek grounds $58/200$ but succeeds on $5/200$; GPT-5.2 grounds $50/200$ but succeeds on $6/200$. These are often not hallucinations, but true, sourced, and thematically nearby statements that do not discharge the proof obligation. The missing capability is assumption accounting: checking objects, domains, hypotheses, parameter ranges, quantifiers, and conclusion strength against the local proof state.

Proof intent must survive a keyword-oriented search interface. Understanding the proof need is not equivalent to retrieving the right document. The shared scholarly backend is closer to keyword and bibliographic matching than to semantic theorem retrieval. A model must therefore express the proof obligation in terms the index can surface. The Steinberg gluing case shows that even natural phrases such as "Steinberg group polynomial gluing property" may fail to surface a usable source; the Muskat case shows a correct compactness need collapsing into generic "Aubin–Lions" search terms. Appendix F.7 gives finer lexicalization probes.

Search helps, but source use remains hard. Table 2 removes open-world source acquisition for materialized author sources, yet the best oracle score is only 12.7% and several models remain below 3%. Thus retrieval is not the only bottleneck: after a paper is found, the model must still locate the theorem boundary and verify that its assumptions and strength fit the proof transition.

Citation-invariant scoring is mathematically necessary. Most successful predictions do not recover the stored citation: Claude succeeds through alternative sources in $13/14$ judge-scored cases, GPT-5.2 in $5/6$, and Kimi, Grok, DeepSeek, and Qwen in all of their successful cases. A citation-locked metric would mark many source-grounded tools wrong, yet alternative-source ToolAcc is still at most 6.5%. A single-annotator audit in Appendix D.4 validates $39/42$ alternative-source successes and identifies three fixed-judge false positives; applying those corrections would only lower the reported scores, so the failure conclusion is unchanged.

Synthesis. The dominant failures are two interfaces: proof obligations must become searchable lexical evidence, and sourced theorems must become assumption-checked proof tools. This motivates theorem-aware retrieval backends that index statement boundaries, assumptions, and conclusion structure under the same frozen-artifact contract.

7. Conclusion

We introduced Re^{2}Math, a benchmark for source-grounded mathematical tool use from research-proof context. Under the fixed-budget shared-backend setting, the strongest fixed-judge result on Eval-200 is only 7.0% ToolAcc, even though several models ground many more statements in real sources. This reveals a central gap in current LLM-based research assistance: models can often find mathematics, but they cannot reliably make it usable. The missing capability is the disciplined composition of retrieval, provenance, theorem localization, and assumption checking into a proof-valid source–tool pair. Re^{2}Math turns this capability into a reproducible evaluation target and points toward theorem-aware retrieval systems that index statements, hypotheses, and conclusion structure rather than paper-level relevance alone.

Broader Impact

Re^{2}Math is intended to support safer and more useful mathematical research assistance. A system that can retrieve source-grounded tools may help researchers find relevant prior work, avoid missing citations, and check whether a proof step relies on assumptions that are actually available in the literature.

The same capability also has risks. If users over-trust model

Table 1. **Core end-to-end results on Eval-200.** Values are percentages with successful counts in parentheses; all denominators are 200. Models are sorted by the primary metric, ToolAcc. Bold marks the best value and underline marks the second best.

	Planning	Citation-side	Grounding	End-to-end
Model	AnchorAcc ↑	CiteRecall@20 ↑	GroundRate ↑	ToolAcc ↑
Claude Opus 4.5	24.0% (48)	**10.5%** (21)	24.0% (48)	**7.0%** (14)
Grok 4	26.5% (53)	4.5% (9)	**41.5%** (83)	<u>3.5%</u> (7)
Kimi K2 Thinking	**43.5%** (87)	6.0% (12)	24.0% (48)	<u>3.5%</u> (7)
GPT-5.2	5.5% (11)	<u>8.5%</u> (17)	25.0% (50)	3.0% (6)
DeepSeek V3.2	8.0% (16)	5.5% (11)	<u>29.0%</u> (58)	2.5% (5)
Gemini 3.1 Pro	2.0% (4)	5.5% (11)	12.0% (24)	2.0% (4)
Qwen3-235B Thinking	<u>40.5%</u> (81)	6.0% (12)	17.0% (34)	1.0% (2)

Table 2. **Oracle-source and citation-invariant diagnostics.** Oracle-source evaluation provides the author-cited source when it can be materialized, isolating within-source localization and extraction. Alternative-source metrics count valid successes through non-cited sources. Oracle values are upper-bound-style diagnostics from the completed oracle runs, not perfectly controlled same-subset reruns.

	End-to-end	Author-source oracle			Citation-invariant success	
Model	ToolAcc	Coverage	Oracle ToolAcc	Δ (pp)	Alt. source ToolAcc	Alt. share among successes
Claude Opus 4.5	**7.0%** (14/200)	35.5% (71/200)	**12.7%** (9/71)	+5.7	**6.5%** (13/200)	92.9% (13/14)
Grok 4	<u>3.5%</u> (7/200)	37.0% (74/200)	2.7% (2/74)	-0.8	<u>3.5%</u> (7/200)	100.0% (7/7)
Kimi K2 Thinking	<u>3.5%</u> (7/200)	37.0% (74/200)	1.4% (1/74)	-2.1	<u>3.5%</u> (7/200)	100.0% (7/7)
GPT-5.2	3.0% (6/200)	35.5% (71/200)	7.0% (5/71)	+4.0	2.5% (5/200)	83.3% (5/6)
DeepSeek V3.2	2.5% (5/200)	37.0% (74/200)	2.7% (2/74)	+0.2	2.5% (5/200)	100.0% (5/5)
Gemini 3.1 Pro	2.0% (4/200)	33.5% (67/200)	<u>7.5%</u> (5/67)	+5.5	1.5% (3/200)	75.0% (3/4)
Qwen3-235B Thinking	1.0% (2/200)	37.0% (74/200)	2.7% (2/74)	+1.7	1.0% (2/200)	100.0% (2/2)

Remark. Coverage is a release-protocol materialization property, not a model score. Main values are fixed-judge benchmark scores. Appendix D.4 reports a single-annotator audit sensitivity analysis identifying three alternative-source false positives; retroactive correction only lowers the affected ToolAcc and alternative-source counts.

outputs, an assistant may propagate incorrect attributions, overstate the applicability of a theorem, or present a weakly related source as sufficient. The benchmark therefore emphasizes grounding and sufficiency, and its intended use is as a diagnostic for human-in-the-loop research support rather than as an automatic substitute for scholarly judgment.

There are also access and representation concerns. A retrieval benchmark built from arXiv-backed published papers will naturally favor English-language, openly available, and source-extractable literature. Fixed-budget offline retrieval lowers compute cost relative to unrestricted browsing agents, but it does not remove the broader dependence on search infrastructure and public-index coverage. Future releases should document domain coverage, source accessibility, and any systematic exclusions introduced by the construction pipeline.

References

Alemi, A. A., Chollet, F., Een, N., Irving, G., Szegedy, C., and Urban, J. Deepmath-deep sequence models for premise selection. *arXiv preprint arXiv:1606.04442*, 2016.

Asai, A., He, J., Shao, R., Shi, W., Singh, A., Chang, J. C., Lo, K., Soldaini, L., Feldman, S., D'arcy, M., Wadden, D., Latzke, M., Tian, M., Ji, P., Liu, S., Tong, H., Wu, B., Xiong, Y., Zettlemoyer, L., Neubig, G., Weld, D., Downey, D., tau Yih, W., Koh, P. W., and Hajishirzi, H. Openscholar: Synthesizing scientific literature with retrieval-augmented lms, 2024. URL https://arxiv.org/abs/2411.14199.

Azerbayev, Z., Piotrowski, B., Schoelkopf, H., Ayers, E. W., Radev, D., and Avigad, J. Proofnet: Autoformalizing and formally proving undergraduate-level mathematics. *arXiv preprint arXiv:2302.12433*, 2023.

Bansal, K., Loos, S., Rabe, M., Szegedy, C., and Wilcox, S. Holist: An environment for machine learning of higher order logic theorem proving. In *International Conference on Machine Learning*, pp. 454–463. PMLR, 2019.

Chen, W., Yin, M., Ku, M., Lu, P., Wan, Y., Ma, X., Xu, J., Wang, X., and Xia, T. Theoremqa: A theorem-driven question answering dataset. *arXiv preprint arXiv:2305.12524*, 2023.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Cohan, A., Feldman, S., Beltagy, I., Downey, D., and Weld, D. S. Specter: Document-level representation learning using citation-informed transformers, 2020. URL https://arxiv.org/abs/2004.07180.

FutureSearch, :, Bosse, N. I., Evans, J., Gambee, R. G., Hnyk, D., Mühlbacher, P., Phillips, L., Schwarz, D., and Wildman, J. Deep research bench: Evaluating ai web research agents, 2025. URL https://arxiv.org/abs/2506.06287.

Färber, M. and Jatowt, A. Citation recommendation: approaches and datasets. *International Journal on Digital Libraries*, 21(4):375–405, August 2020. ISSN 1432-1300. doi: 10.1007/s00799-020-00288-2. URL http://dx.doi.org/10.1007/s00799-020-00288-2.

Gao, B., Song, F., Yang, Z., Cai, Z., Miao, Y., Dong, Q., Li, L., Ma, C., Chen, L., Xu, R., et al. Omni-math: A universal olympiad level mathematic benchmark for large language models. *arXiv preprint arXiv:2410.07985*, 2024a.

Gao, G., Ju, H., Jiang, J., Qin, Z., and Dong, B. A semantic search engine for mathlib4. *arXiv preprint arXiv:2403.13310*, 2024b.

Gao, T., Yen, H., Yu, J., and Chen, D. Enabling large language models to generate text with citations, 2023. URL https://arxiv.org/abs/2305.14627.

Glazer, E., Erdil, E., Besiroglu, T., Chicharro, D., Chen, E., Gunning, A., Olsson, C. F., Denain, J.-S., Ho, A., Santos, E. d. O., et al. Frontiermath: A benchmark for evaluating advanced mathematical reasoning in ai. *arXiv preprint arXiv:2411.04872*, 2024.

Google AI for Developers. Gemini deep research agent. https://ai.google.dev/gemini-api/docs/deep-research, 2026.

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

He, C., Luo, R., Bai, Y., Hu, S., Thai, Z., Shen, J., Hu, J., Han, X., Huang, Y., Zhang, Y., et al. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3828–3850, 2024.

He, Q., Pei, J., Kifer, D., Mitra, P., and Giles, C. L. Context-aware citation recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, pp. 421–430, 2010. doi: 10.1145/1772690.1772734.

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., and Steinhardt, J. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

Hubert, T., Mehta, R., Sartran, L., Horváth, M. Z., Žužić, G., Wieser, E., Huang, A., Schrittwieser, J., Schroecker, Y., Masoom, H., et al. Olympiad-level formal mathematical reasoning with reinforcement learning. *Nature*, pp. 1–3, 2025.

Kaliszyk, C., Chollet, F., and Szegedy, C. Holstep: A machine learning dataset for higher-order logic theorem proving. *arXiv preprint arXiv:1703.00426*, 2017.

Li, M., Zeng, Y., Cheng, Z., Ma, C., and Jia, K. Reportbench: Evaluating deep research agents via academic survey tasks, 2025. URL https://arxiv.org/abs/2508.15804.

Lála, J., O'Donoghue, O., Shtedritski, A., Cox, S., Rodriques, S. G., and White, A. D. Paperqa: Retrieval-augmented generative agent for scientific research, 2023. URL https://arxiv.org/abs/2312.07559.

Mahdavi, S., Li, M., Liu, K., Thrampoulidis, C., Sigal, L., and Liao, R. Leveraging online olympiad-level math problems for llms training and contamination-resistant evaluation. *arXiv preprint arXiv:2501.14275*, 2025.

Mikuła, M., Tworkowski, S., Antoniak, S., Piotrowski, B., Jiang, A. Q., Zhou, J. P., Szegedy, C., Kuciński, Ł., Miłoś, P., and Wu, Y. Magnushammer: A transformer-based approach to premise selection. *arXiv preprint arXiv:2303.04488*, 2023.

OpenAI. Learning to reason with LLMs. https://openai.com/index/learning-to-reason-with-llms/, 2024.

OpenAI. Introducing deep research. https://openai.com/index/introducing-deep-research/, 2025.

Petroni, F., Piktus, A., Fan, A., Lewis, P., Yazdani, M., Cao, N. D., Thorne, J., Jernite, Y., Karpukhin, V., Maillard, J., Plachouras, V., Rocktäschel, T., and Riedel, S. Kilt: a benchmark for knowledge intensive language tasks, 2021. URL https://arxiv.org/abs/2009.02252.

Sanderson, M. Test collection based evaluation of information retrieval systems. *Foundations and Trends in Information Retrieval*, 4(4):247–375, 2010. doi: 10.1561/1500000009.

Singh, A., D'Arcy, M., Cohan, A., Downey, D., and Feldman, S. Scirepeval: A multi-format benchmark for scientific document representations, 2023. URL https://arxiv.org/abs/2211.13308.

Sun, H., Min, Y., Chen, Z., Zhao, W. X., Fang, L., Liu, Z., Wang, Z., and Wen, J.-R. Challenging the boundaries of reasoning: An olympiad-level math benchmark for large language models. *arXiv preprint arXiv:2503.21380*, 2025.

Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., and Gurevych, I. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models, 2021. URL https://arxiv.org/abs/2104.08663.

Tsoukalas, G., Lee, J., Jennings, J., Xin, J., Ding, M., Jennings, M., Thakur, A., and Chaudhuri, S. Putnambench: Evaluating neural theorem-provers on the putnam mathematical competition. *Advances in Neural Information Processing Systems*, 37:11545–11569, 2024.

Urban, J. Mptp 0.2: Design, implementation, and initial experiments. *Journal of Automated Reasoning*, 37(1): 21–43, 2006.

Wadden, D., Lin, S., Lo, K., Wang, L. L., van Zuylen, M., Cohan, A., and Hajishirzi, H. Fact or fiction: Verifying scientific claims, 2020. URL https://arxiv.org/abs/2004.14974.

Wadden, D., Lo, K., Kuehl, B., Cohan, A., Beltagy, I., Wang, L. L., and Hajishirzi, H. Scifact-open: Towards open-domain scientific claim verification, 2022. URL https://arxiv.org/abs/2210.13777.

Wang, P., Li, L., Shao, Z., Xu, R., Dai, D., Li, Y., Chen, D., Wu, Y., and Sui, Z. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439, 2024.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Welleck, S., Liu, J., Bras, R. L., Hajishirzi, H., Choi, Y., and Cho, K. Naturalproofs: Mathematical theorem proving in natural language. *arXiv preprint arXiv:2104.01112*, 2021.

White, C., Dooley, S., Roberts, M., Pal, A., Feuer, B., Jain, S., Shwartz-Ziv, R., Jain, N., Saifullah, K., Naidu, S., et al. Livebench: A challenging, contamination-free llm benchmark. *arXiv preprint arXiv:2406.19314*, 4, 2024.

Xin, H., Ren, Z., Song, J., Shao, Z., Zhao, W., Wang, H., Liu, B., Zhang, L., Lu, X., Du, Q., et al. Deepseek-prover-v1. 5: Harnessing proof assistant feedback for reinforcement learning and monte-carlo tree search. *arXiv preprint arXiv:2408.08152*, 2024.

Yang, K., Swope, A., Gu, A., Chalamala, R., Song, P., Yu, S., Godil, S., Prenger, R. J., and Anandkumar, A. Leandojo: Theorem proving with retrieval-augmented language models. *Advances in Neural Information Processing Systems*, 36:21573–21612, 2023.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., and Narasimhan, K. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.

Yue, A. S., Madaan, L., Moskovitz, T., Strouse, D., and Singh, A. K. Harp: A challenging human-annotated math reasoning benchmark. *arXiv preprint arXiv:2412.08819*, 2024.

Zheng, K., Han, J. M., and Polu, S. Minif2f: a cross-system benchmark for formal olympiad-level mathematics. *arXiv preprint arXiv:2109.00110*, 2021.

A. Additional Discussion: What the Benchmark Is Designed to Reveal

This section discusses the methodological implications of Re^{2}Math independently of any specific model ranking. The goal is to clarify what kinds of capability gaps the benchmark is intended to expose, and why the current evaluation is organized around source-grounded success rather than citation recovery alone.

A.1. Author Citations Provide Supervision, Not the Definition of Success

Instrumental citations are a strong construction signal because they tell us that a human author relied on an external result at a specific proof transition. But the cited source is often not the only valid way to justify the step. The same theorem may appear in multiple papers, books, surveys, or lecture notes, and a stronger statement from a different source may also be perfectly usable.

This is why Re^{2}Math is citation-invariant but source-grounded. A citation-locked benchmark would underestimate genuine literature competence by penalizing alternative valid sources. By contrast, our official success criterion requires only that the returned statement be grounded in the selected source and sufficient for the gap.

A.2. Why We Separate Planning, Retrieval, and Within-Source Extraction

Mathematical tool use is a pipeline, not a single decision. A system may fail because it cannot infer the right kind of next step from proof context, because it cannot turn that need into a useful search query, because it cannot recognize the right source from metadata, or because it cannot localize the correct statement inside the selected document.

The benchmark therefore fixes separate views of the problem. Raw-track ablations ask whether the hierarchical context supports the intended proof-level planning signal. End-to-end evaluation measures actual open-world retrieval and grounded extraction. Oracle-source evaluation then removes open-world retrieval and isolates within-source localization and sufficiency. This decomposition is designed to make future method improvements more attributable.

A.3. Why Source-Level Failure Decomposition Matters

A single low end-to-end score is hard to interpret. It does not tell us whether the final output lacked any source support, whether it was grounded but mathematically insufficient, or whether the source–tool pair succeeded. Because Re^{2}Math is citation-invariant, the paper-facing failure decomposition is source-level rather than citation-conditioned: it partitions outputs by GROUND, SUFF, and TOOLACC, not by whether the author-cited source appears in Top-20.

This distinction is especially important in mathematics. A model may retrieve a thematically related source and extract a true theorem, yet the theorem may have incompatible assumptions, a weaker conclusion, or the wrong objects. Retrieval-conditioned views based on $\mathrm{CiteRecall@20}$ remain useful for debugging, but they are not the official failure definition.

A.4. Deployment Boundary

Even a strong score on Re^{2}Math would not make a system an autonomous proof assistant in the formal sense. The benchmark evaluates source-grounded tool use under natural-language proof context, not mechanically checked proof completion. In practical workflows, the most immediate use case is therefore a research copilot that helps surface candidate tool families, likely sources, and plausible theorem statements for human verification.

At the same time, Re^{2}Math offers a clearer target for future systems that combine retrieval, verification, and formalization. Because the task explicitly requires source grounding and applicability under context, it can serve as a bridge between open-world mathematical search and downstream formal proof checking.

B. Dataset Construction Protocol

This appendix specifies the release protocol behind Re^{2}Math. The goal is to make the benchmark auditable along three axes: (i) how proof-gap instances are mined from published mathematical writing, (ii) how the release and evaluation split are frozen, and (iii) how model outputs are rendered, retrieved, judged, and aggregated. We use the notation of Section 3; implementation-specific choices are described here rather than in the task definition.

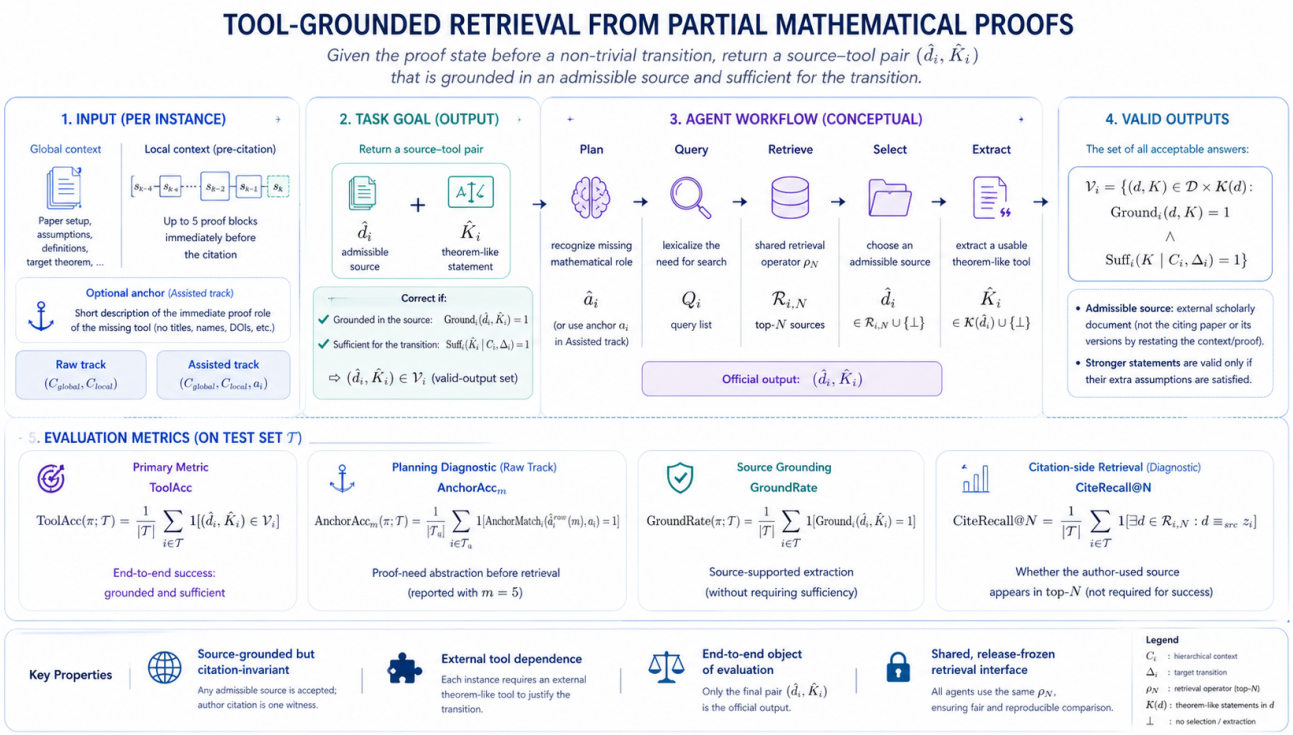

Figure 2. **Overview of tool-grounded retrieval.** Given hierarchical proof context (global + local) and an optional anchor, the agent produces a source–tool pair via planning, query formulation, retrieval, source selection, and statement extraction. A prediction is correct if the extracted statement is grounded in the selected source and sufficient for the target proof transition. The figure is illustrative; formal definitions are given in Section 3.

Illustrative overview. For readers who prefer a high-level, visual summary, we provide an illustrative diagram of the task pipeline in Figure 2. The figure depicts the end-to-end workflow—from proof context and (optional) anchor, through query formulation and retrieval, to source selection and tool extraction—and highlights the validity criterion based on grounding and proof-gap sufficiency. This diagram is intended for intuition only and does not replace the formal task definition in Section 3, which specifies the precise inputs, outputs, and evaluation criteria.

B.1. Published-First Manifest and arXiv Pairing

Re^{2}Math is constructed under a *published-first, arXiv-backed* policy. Candidate citing papers are selected from the published literature, so each instance inherits a real venue, citation record, and scholarly identity. When a high-confidence arXiv counterpart with LaTeX source is available, we use that source for extraction, alignment, and proof-local parsing. This separation lets the benchmark retain a published-literature identity while avoiding brittle OCR over publisher PDFs.

The manifest records two linked identities for every accepted paper. The *published identity* stores title, authors, venue, year, publication date, DOI, and related bibliographic metadata. The *extraction identity* stores the arXiv identifier, source URL, matched title, match type, and match confidence. Matching proceeds through a conservative cascade: DOI-exact agreement, then high-confidence title–author–year agreement, then manual verification. Ambiguous pairs are sent to a review queue rather than forced into the release.

B.2. Two-Stage Mining Pipeline

Stage I: paper-level proof scouting. The first stage operates at the paper level. Given a matched arXiv source, a compact bibliography map, and a target-theorem hint, the scout extracts three objects: a global setup, the proof span of the target theorem, and proof-internal citation occurrences that appear to function as external mathematical tools. This stage deliberately excludes historical remarks, background citations, definitions, and citations outside the target proof.

Table 3. **Construction protocol overview.** The pipeline separates paper identity, proof-local mining, instance formation, and release curation.

Component	Input	Operation	Release artifact
Manifest	Published mathematical papers	Match to arXiv source when available; preserve both identities	Published–arXiv pair
Paper scouting	Matched LaTeX source and bibliography	Locate setup, target theorem proof, and instrumental proof citations	Proof-local citation candidates
Citation extraction	Citation-local proof slice	Build pre-citation context, anchor, reference witness, and citation metadata	Canonical row
Curation	Candidate rows	Deduplicate, repair, filter, and split at citing-paper level	Frozen release and Eval-200

Stage II: citation-level instance extraction. The second stage processes one retained citation occurrence at a time. It aligns the citation inside a proof-local slice, extracts the ordered pre-citation blocks, and constructs the triple

$$(\mathsf{x}_i, \mathsf{y}_i, \mathsf{z}_i) = \big((\mathcal{C}^i_{\text{global}}, \mathcal{C}^i_{\text{local}}, a_i), K^\star_i, d^{\text{cite}}_i\big)$$

used in the main text. If the citation cannot be localized reliably, the candidate is dropped. We do not backfill missing local context from unrelated parts of the proof.

B.3. Anchor Resolution and Leakage Controls

The anchor hint a_i is a planning cue, not an answer. It is produced by combining an LLM-generated proof-intention description, citation-local heuristics, implication-style rewrites from the locator snippet, and local-context heuristics. The final anchor is designed to be leakage-controlled: it may describe the mathematical role of the missing tool, but it should not include theorem titles, author names, citation keys, source titles, DOI strings, arXiv identifiers, or other bibliographic shortcuts. Manual repairs are applied only to remove leakage, repair malformed text, or make the intended proof move grammatically clear.

B.4. Construction Prompt Templates

The construction prompts below are written as release-facing prompt contracts. The actual released code may call fallback variants depending on parse failures or weak tool statements, but all variants preserve the same principle: recover proof-local external tool use without inventing mathematical content.

C1. Stage-1 structure and citation scouting

Role. Extract one benchmark instance from one mathematical paper. Return JSON only.
Input payload. A compact bibliography map and a compact paper view constructed from the matched arXiv source.
Required fields.

- `global_context.setup`: concise setup needed for the paper's main theorem.
- `global_context.target_theorem`: the main theorem only; return an empty string if uncertain.
- `proof_span.start_snippet` and `proof_span.end_snippet`: verbatim 20–40 word snippets delimiting the proof of the target theorem when recoverable.
- `proof_citations`: at most six citations used as external mathematical tools inside that proof. Each item contains a citation key, locator snippet, and short usage reason.

Rules. Use only citation keys from the bibliography map. Exclude citations used only for background, setup, history, definitions, or non-proof discussion. Prefer empty fields over guessing. Keep the setup and target theorem compact.

C1-fallback. Stage-1 recall and repair prompts

Compact retry. When a reasoning model truncates or fails to emit valid JSON, return one compact JSON object with the same keys as C1, no reasoning, no markdown, and at most four proof citations.
Targeted recall. When the proof span or proof citations are missing, act as a Mathematical Proof Citation Scout. Given a target-theorem hint, optional setup, bibliography map, and compact paper view, recover the proof span and recall plausible

external-tool citations inside that proof. Favor recall over precision, but discard background, historical, definitional, or non-proof citations.
Proof-local recall. When the proof span has already been identified, read only the provided proof text and recover citations used as external theorems, lemmas, propositions, criteria, estimates, inequalities, or structural facts. Return JSON with a `proof_citations` list.
Structured repair. When a prompt output is malformed, repair syntax into the requested JSON schema while preserving content when possible. Do not invent new mathematical content beyond minimal syntax repair.

C2. Stage-2 citation-local witness extraction

Role. Mathematical Logic Expert.
Input payload. Target theorem, target citation key, pre-citation local proof blocks, citation-focus snippet, short tool-usage reason, and bibliography entry.
Task. Recover only the grounded external tool for the given citation occurrence. Focus on the proof step immediately around the citation. Return the minimal theorem-like statement, estimate, implication, or criterion needed at this step.
Rules.

- Do not rewrite the local context blocks.
- Do not output a theorem title alone.
- Prefer an explicit implication, estimate, criterion, or theorem-like statement over vague prose.
- Reject discourse wrappers such as "Recall that," "In addition," case-analysis narration, future-work text, examples, and remarks.
- Set `restated_in_citing_paper=true` only when the citing text explicitly states the mathematical implication or estimate itself.
- The returned statement must not include theorem labels or source-reporting wrappers such as "Proposition 1.5 states that."

Output. JSON with `reference_tool_latex`, `reference_tool_type`, and `restated_in_citing_paper`.

C2-fallback. Focused tool extraction and implication rewrite

Trigger. Used when the default Stage-2 output is empty, generic, wrapped in narrative text, or not theorem-like enough for a benchmark row.
Primary retry. Focus only on recovering the external tool statement used by the specific citation. If the paper only reveals the consequence used at this step, return that consequence as a faithful theorem-like paraphrase. Do not return an empty string unless no recoverable theorem-like content is provided.
Implication pass. If the citation sentence says that an object belongs to one class and therefore belongs to another class by a cited result, rewrite the operative content as a theorem-like implication of the form "If ... then ...". A short usable implication is preferred over an empty answer.
Batch extraction. When enabled, the same fields are extracted for multiple citation candidates jointly; each item still returns a citation key, tool statement, tool type, and restatement flag.

B.5. Curation, Schema, Release, and Split Policy

After mining, the curation step deduplicates candidate rows, repairs broken core fields, removes low-confidence alignments, and verifies that all required task fields are present. The canonical schema follows the x/y/z split in Section 3: x stores the proof context and optional anchor, y stores the reference theorem-like witness, and z stores auxiliary citation metadata.

The local proof context stores at most the last five pre-citation proof blocks in order. Local-window analyses are obtained by slicing this stored list from the end with $m \in \{1, 3, 5\}$. If fewer than m blocks are available, the evaluator uses the shorter available context; if no pre-citation block exists, the local context remains empty rather than fabricated.

Each paper-facing comparison is tied to a frozen release: a fixed instance set, an official held-out evaluation split, retrieval provider and externally controlled settings, query log, and cached candidate metadata. The balanced Eval-200 subset is the official held-out evaluation split used in the main experiments. The remaining curated rows form a non-test pool for future train/dev use. The current paper does not train, tune, or select prompts on Eval-200. Future partitions should be made at the citing-paper level rather than the proof-gap level, because a single citing paper may contribute multiple related gaps.

Eval-200 selection. The held-out evaluation split is quality-filtered rather than randomly sampled. The selection script requires a journal-paper cited source, nonempty locator, nonempty leakage-controlled anchor, nonempty complete reference witness, and nonempty local context. It then ranks candidates by a quality score, enforces a per-paper cap, and fills a

Table 4. **Release and leaderboard contract.** This table records the artifact components that make a living benchmark comparable across submissions.

Artifact	Frozen contents	Comparison rule
Dataset release	Instance set, schema, split, domain labels, construction manifest	Results are tied to a named release.
Retrieval artifact	Retrieval provider and externally controlled settings, evaluated queries, Top-20 candidate metadata, timestamps, aggregation rules	New query logs require a new frozen artifact or re-evaluation.
Source materialization log	Resolved source identifiers, full-text access status, parsing status	Oracle coverage is interpreted as materialization coverage, not model accuracy.
Scoring artifact	Judge version, rubric prompts, source matcher, aggregation script	Scores should be recomputed if the judge or matcher changes.

Table 5. **Single-annotator construction audit.** The audit covers 120 independent retained instances and is used to characterize construction risk, not to rescore Eval-200. It validates local alignment, reference-witness sufficiency, metadata, and external-source dependence, while identifying citation-role instrumentality and anchor sanitization as the main residual construction risks.

Audit item	Pass condition	Positive / audited	Rate
Instrumental citation	Citation supplies a non-trivial external tool for the local proof transition	110 / 120	91.7%
Local-context alignment	Stored pre-citation blocks are the exposed proof state for the target citation	120 / 120	100.0%
Reference witness sufficiency	Curated $K_i^\star$ is sufficient for the gap under the stored context	120 / 120	100.0%
Anchor leakage control	Anchor contains no theorem title, author, citation key, DOI, arXiv ID, or source title	98 / 120	81.7%
Citation metadata correctness	Stored d_i^{cite} matches the source used by the citing paper	116 / 120	96.7%
External-source constraint	Instance requires an external tool rather than merely restating the citing proof	120 / 120	100.0%

balanced number of instances in each benchmark domain.

Leaderboard and release policy. A public comparison should identify the full frozen tuple: dataset release, evaluation split, retrieval provider and externally controlled settings, retrieval/query-cache artifact, source-materialization log, scoring artifact, and evaluation date. A future model may generate queries that are absent from an earlier cache; in that case, the benchmark records a new query-result artifact under the same retrieval provider and externally controlled settings rather than silently mixing live search with cached results. Results from different dataset releases or retrieval artifacts should not be treated as entries on the same leaderboard unless older systems are re-evaluated under the same artifact.

Anonymous release access. An anonymous reviewer-accessible release is available at https://github.com/Anonymous1-bot-bit/Re-2Math. The release contains the curated instances, Eval-200 split, non-test pool, schema, retrieval/query-cache artifact, source-materialization log, scoring artifact, prompt contracts, aggregation scripts, and Croissant metadata. The release will be made public and de-anonymized by the camera-ready deadline if accepted.

B.6. Construction Quality Audit

Because the construction pipeline is automated and LLM-assisted, we audited a fixed subset of 120 independent retained instances. The audit checks whether each row is genuinely proof-local, externally tool-dependent, and leakage-controlled. This is a single-annotator construction audit rather than an inter-annotator reliability study. Table 5 summarizes the results.

The two weakest dimensions are citation-role instrumentality and anchor sanitization. We interpret these flags conservatively. Instrumentality is a construction-signal audit: a failure means that the original citation is not a clean witness for the local proof transition, not that the benchmark row has no valid source–tool solution. In the same audit, all 120 rows passed reference-witness sufficiency and the external-source constraint, so the scored task remains the final source–tool predicate rather than exact recovery of the author citation.

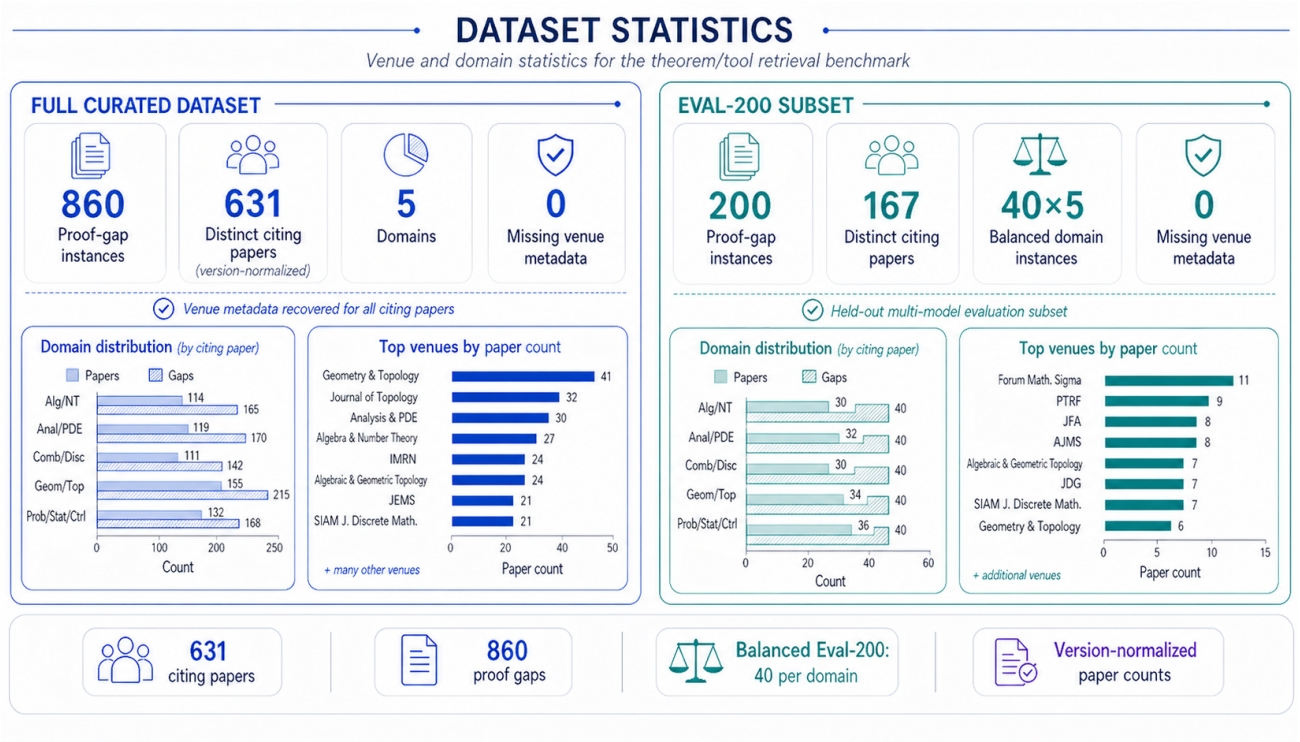

Figure 3. **Dataset statistics.** Overview of the curated benchmark and the Eval-200 subset. The full curated dataset contains 860 proof-gap instances from 631 version-normalized citing papers. Eval-200 contains 200 held-out proof-gap instances from 167 citing papers, balanced across five domains.

There were 10 instrumentality flags. In a triaged subset of five flagged cases, the issues consisted of background/historical citation roles (3) and comparison/related-work citation roles (2); no pure definition-citation or locally self-contained proof-step case was observed in that triage. The remaining flagged cases are therefore treated conservatively as citation-role ambiguity. Under the frozen-release contract, post-hoc removal of rows after evaluation would define a new dataset release and require corresponding retrieval and scoring artifacts. We therefore report these flags as residual construction noise in the author-citation witness and use them to tighten future citation filters, rather than changing the fixed Eval-200 denominator post hoc.

Anchor-sanitization failures are handled separately because anchors are not exposed in the Raw end-to-end leaderboard. They can affect Assisted-track query diagnostics and may make stored anchors imperfect as planning targets; accordingly, assisted-query results are reported only as diagnostics and not as official task scores. The validity predicates Ground and Suff are evaluated on the model's selected source–tool pair and do not use anchor text as evidence.

B.7. Snapshot Statistics

The main text reports compact domain coverage. Figure 3 gives the full release summary used for sanity checks and reproducibility.

C. Evaluation Protocol and Prompt Templates

This section gives the implementation of the abstract retrieval operator and validity predicates used in the main text. The primary condition disables model-native web-search, browser, file-search, and deep-research tools. Models may generate queries, but query execution and candidate retrieval are supplied by the frozen Google Scholar retrieval artifact.

C.1. Input Rendering and Retrieval Artifact

Input rendering. The evaluator renders inputs directly from canonical rows. On the Raw track, a model receives the global setup, target theorem, and ordered local proof blocks. On the Assisted track, the same context is shown with the stored sanitized anchor hint. The canonical end-to-end condition uses the Raw track with the five-block local window.

Why Google Scholar. The paper-facing retrieval artifact is instantiated with Google Scholar because the task targets open-world mathematical tool discovery across heterogeneous scholarly sources. The needed tool may appear in a published article, an arXiv or other preprint record, a proceedings paper, a book, a survey, or a bibliographic record that links to a resolvable source. An arXiv-only index would make construction easier to reproduce but would bias evaluation toward recent and source-available papers; a publisher-specific index would miss many preprints, books, and older mathematical records. Google Scholar provides a broad scholarly-search interface across disciplines and source types. We therefore use it as a standardized discovery layer, while freezing its observed outputs because its internal corpus coverage and ranking function are not fully specified by us.

Retrieval execution and temporal controls. The paper-facing retrieval artifact is a frozen Google Scholar query cache. It is not defined by a common start date. Instead, retrieval uses a per-instance bibliographic cutoff. For each instance i, the evaluator runs model-authored queries with no lower-bound year and with an upper-bound year equal to the publication year of the original citing paper. This policy is intended to prevent later literature from being used to solve a proof gap from an earlier paper, while preserving access to older papers, books, surveys, and preprints. Because Google Scholar filtering is year-level, this is a publication-year cutoff rather than a reconstruction of the exact day-level search state available to the original authors.

For each model-authored query, the evaluator executes the query once through Google Scholar under fixed externally controlled settings and caches the returned Top-20 candidate prefix. The fixed settings were:

interface/result language	= English,
region or domain	= not additionally controlled,
ranking mode	= default relevance ranking,
date-range policy	= no lower-bound year; upper year = citing-paper publication year,
patent results	= excluded.

The cache records the query string, execution timestamp, returned rank order, candidate metadata, the externally controlled Google Scholar settings, and source-materialization outcomes when materialization is attempted. Reported numbers are computed only from this frozen query-result artifact.

Temporal interpretation. The bibliographic cutoff and the artifact freeze serve different purposes. The per-instance Google Scholar date filter restricts candidate records to the literature dated no later than the original citing paper's publication year. The artifact freeze makes the observed results reproducible: later changes in Google Scholar coverage, metadata, or ranking are not incorporated into the reported numbers. We do not manually promote the author-cited source, rerank candidates using the stored witness, or expose hidden retrieval signals to models. Accordingly, $\mathrm{CiteRecall@20}$ should be interpreted as a diagnostic of model query choice interacting with the frozen Google Scholar artifact under this date-filter policy, not as a reconstruction of the exact historical search process of the original authors.

Candidate composition. Returned candidates may include journal or proceedings articles, arXiv or other preprint records, books, surveys, publisher pages, and bibliographic records when surfaced by Google Scholar. Candidate type alone does not determine correctness: a source is admissible only if the model's extracted statement is grounded in that source and sufficient for the proof transition. Conversely, the author-cited source is used for construction and diagnostics, but exact recovery of that source is not required for $\mathrm{ToolAcc}$.

Ranking. The rank order used in the benchmark is the order returned by Google Scholar for the cached query under the fixed settings above. We do not manually promote the author-cited source, rerank candidates using the stored witness, or expose hidden retrieval signals to models. Thus ρ_N in the task definition should be understood as the frozen observed mapping from model-authored queries to cached ranked candidate prefixes under this release artifact.

Source materialization. After inspecting Top-20 metadata, the model may select at most one candidate to open. If the selected candidate can be resolved under the release protocol, the solver receives the fixed text materialization used by all models. This materialization step deliberately avoids comparing model-native PDF readers, OCR systems, equation renderers, or proprietary document-ingestion tools in the primary leaderboard. The source-materialization log records whether a selected or oracle source was resolved successfully.

Interpretation of retrieval diagnostics. $\mathrm{CiteRecall@20}$ asks whether the stored author-cited witness appears among the cached Top-20 candidates under the source-identity matcher in Section C.2. It is therefore a diagnostic of the interaction between model query choice, Google Scholar coverage, Google Scholar ranking, and source-identity matching. It is not a pure model-ability score and is not the official success criterion. The official criterion remains citation-invariant: any admissible grounded source–tool pair can receive credit.

C.2. Source Identity Matching

Citation-side diagnostics use a source-identity matcher $\mathsf{Match}(d, \mathsf{z}_i)$ between retrieved candidate metadata and stored citation metadata. The matcher prioritizes DOI and arXiv identifiers when available; otherwise it applies normalized title–author matching, with ambiguous journal/preprint variants handled by conservative or manual rules. This matcher is used only for $\mathrm{CiteRecall@}N$ and oracle-source diagnostics. It is not the task definition: correctness is determined by Ground, Suff, and membership in $\mathcal{V}_i$.

Citing-paper exclusion. The citing paper itself, including its arXiv and published variants, is not an admissible retrieved source for the target gap merely because it contains the citation context. A prediction that restates the local proof sentence or points back to the citing proof does not satisfy Ground. The only exception is when the selected text independently states or proves the external tool outside the local citation-dependent argument; otherwise the benchmark would collapse into finding the original paper rather than retrieving the invoked scholarly tool.

C.3. Source Shortlisting and Full-Text Extraction

The model does not read all retrieved documents. It first shortlists from Top-20 metadata, then opens at most one selected source in full text when that source can be materialized. The solver extracts, or faithfully restates, a theorem-like statement from the selected source. A prediction can therefore fail at source selection, at grounding, or at sufficiency even when an apparently relevant paper appears in the candidate list.

C.4. Evaluation Prompt Templates

The following prompt contracts are complete with respect to role, input, constraints, and output fields. Full released prompt strings preserve these contracts while adding instance renderings and formatting delimiters.

E1. Planning prompt

Role. Mathematical Research Assistant.
Input. Global context, target theorem, and pre-citation local proof window.
Task. Infer the immediate next proof intention before retrieval. Describe what external theorem, estimate, criterion, structural result, comparison principle, gluing property, compactness result, or inequality would close the next proof transition.
Constraint. Return a sanitized planning hint. Do not mention theorem titles, author names, citation keys, source titles, DOI strings, arXiv identifiers, or bibliographic metadata.
Output. JSON with field `planning_anchor`.

C.5. AnchorMatch Rubric and Examples

Purpose. $\mathrm{AnchorAcc}$ is a Raw-track planning diagnostic, not an end-to-end success metric. It asks whether the model can infer the immediate proof need before retrieval. The stored anchor is a sanitized description of the intended proof role of the missing external tool; it is not a theorem title, source identifier, or gold answer. Accordingly, $\mathsf{AnchorMatch}$ compares proof intentions rather than bibliographic identities.

Rubric. A model-produced planning description matches the stored anchor when it identifies the same local mathematical obligation: the relevant objects, the direction of use, and the type of conclusion needed to continue the proof. Exact wording

is not required. Benign paraphrases, equivalent terminology, or a different theorem family can match if they describe the same proof role. By contrast, broad topical overlap is insufficient. A description that only says "use compactness," "apply a comparison theorem," or "find a result about positive definite functions" does not match unless it specifies the objects and the proof obligation tightly enough to guide retrieval. Bibliographic shortcuts are also insufficient: mentioning an author, theorem name, source title, citation key, DOI, or arXiv identifier does not by itself establish an anchor match.

Relation to sufficiency. Anchor matching is weaker than Suff. It does not require the model to state a theorem with all hypotheses, and it does not verify that any source actually contains the needed result. It only evaluates whether the model has formed the right pre-retrieval intention. A model can therefore have high $\mathrm{AnchorAcc}$ but low $\mathrm{ToolAcc}$ if its queries fail, if it selects the wrong source, or if it extracts a grounded but insufficient statement. Conversely, a model can occasionally fail $\mathrm{AnchorAcc}$ but still succeed end-to-end by finding a valid alternative route.

Examples. A positive match requires the same proof role, not the same wording. For example, if the stored anchor asks for a compactness result that turns uniform bounds and time-translation control into strong convergence in the topology needed for a nonlinear limit passage, then a model description invoking an Aubin–Lions or Rellich-type compactness theorem for that specific convergence role is counted as a match. Similarly, if the stored anchor asks for a local-to-global gluing principle for a group-theoretic construction, then a model description of a patching theorem upgrading locally verified data to the desired global equality is counted as a match.

Negative cases are also important. A response such as "use harmonic analysis on spheres" does not match an anchor asking for a characterization of positive definite functions on spheres via expansion coefficients, because it is only topically related. A response asking for a controllability theorem does not match an anchor asking for non-controllability in a specified fractional-heat parameter regime, because the direction of use is wrong. A bibliographic shortcut such as "find the theorem in the cited paper" also does not match unless it states the mathematical proof role. Borderline descriptions such as "use compactness to get a subsequence" are marked uncertain or negative when they omit the topology, objects, or conclusion needed by the local proof step; only clear positive matches contribute to $\mathrm{AnchorAcc}$.

E2. Query-generation prompt

Role. Mathematical literature search assistant.
Input. Global setup, target theorem, local proof window, and optionally a planning anchor.
Task. Write a concise scholarly search query aimed at finding a theorem, estimate, lemma, criterion, or standard result that can justify the next proof step. Prefer mathematical objects, canonical terminology, named properties, theorem families, and strength indicators over generic application words.
Rules. Avoid copying unique prose from the citing paper. Do not use author names, citation keys, DOI strings, arXiv identifiers, or bibliographic metadata from the stored witness. Keep the query short enough for a keyword-oriented scholarly search backend.
Output. JSON with field `query`.

E3. Metadata shortlisting prompt

Role. Careful mathematical retrieval shortlister.
Input. Proof context, planned proof need, search query, and Top-k candidate metadata.
Task. Select the single candidate source most worth opening in full text. Decide from title, authors, venue, snippets, and metadata whether the candidate is likely to contain a theorem-like tool for the proof gap.
Rules. Do not solve the proof from metadata alone. Do not select the citing paper merely because it contains the citation context. Prefer sources likely to state or prove the external tool.
Output. JSON with selected candidate index, selected title, and a brief reason.

E4. Full-text solver prompt

Role. Mathematical assistant for the gap-filling task.
Input. Benchmark context and the full text of one selected candidate paper, truncated to a fixed prompt budget.
Task. Decide whether the selected paper contains a theorem-like statement that bridges the proof gap. If so, extract or faithfully restate the minimal relevant theorem, lemma, proposition, corollary, criterion, estimate, or standard result.
Rules. Preserve hypotheses, domains, quantifiers, parameter ranges, and conclusion strength. Do not return a theorem title alone. Do not use the local citing proof sentence as the source of the tool. If the paper does not support a usable statement, abstain.
Output. JSON with `relevance_score`, short reason, and `extracted_theorem`.

Table 6. **Oracle-source materialization summary.** Counts describe the release-protocol source-materialization artifact, not model capability.

Materialization status	Count	Interpretation
Unified oracle-evaluable subset	74/200	Author-cited source could be resolved and materialized in the release-protocol artifact.
Materialized in every completed run	66/200	Stable subset shared across all completed oracle runs.
Unavailable in every completed run	126/200	Source could not be materialized under the current release protocol.
Run-specific materialization differences	8/200	Cases with materialization/logging differences across completed oracle runs.

E5. Grounding-judge prompt

Input. Selected source identifier, predicted theorem-like statement, and source excerpts selected by overlap with the prediction.
Task. Return `yes`, `no`, or `uncertain` depending on whether the predicted statement is verbatim or a mathematically faithful restatement of a theorem-like statement supported by the selected source.
Scope. Grounding verifies provenance, not independent truth of the source theorem. A statement can be grounded even if it is not sufficient for the proof gap.
Output. Label, short reason, confidence, and evidence span.

E6. Sufficiency-judge prompt

Input. Rendered proof context, reference witness $K^\star$, and predicted theorem statement $\hat{K}$.
Task. Return `yes`, `no`, or `uncertain` depending on whether $\hat{K}$, together with the proof context, is sufficient to justify the target proof transition. Use $K^\star$ as one reference witness for the intended proof role, but do not require exact identity with $K^\star$. Accept equivalent, stronger, or otherwise admissible alternative statements when their assumptions are established by the context and their conclusions close the same gap.
Checks. Verify objects, domains, hypotheses, parameter ranges, quantifiers, and conclusion strength. Penalize true but weaker, mismatched, or inapplicable statements.
Output. Label, short reason, and confidence.

E7. Oracle-source extraction prompt

Role. Mathematical assistant running an oracle-source extraction experiment.
Input. Benchmark context and the materialized author-cited source content.
Task. Because the source is guaranteed to contain the author-used witness when materialized, identify and extract the theorem-like statement that solves the gap.
Output. JSON with short reason and `extracted_theorem`.

C.6. Grounding, Sufficiency, and Oracle Evaluation

A held-out judge scores two predicates. GROUND checks whether the predicted theorem-like statement is supported by the selected source; it verifies provenance, not the mathematical truth of the source theorem itself. SUFF checks whether the statement closes the proof gap under the given context. TOOLACC is their conjunction. Oracle-source evaluation bypasses open-world retrieval: when the cited source can be resolved and materialized, the model is given that source directly and asked to extract the relevant theorem-like statement. This diagnostic isolates within-source localization, statement extraction, and applicability checking from candidate acquisition.

C.7. Oracle Source Materialization

Oracle-source evaluation is defined only for instances whose author-cited source can be resolved and materialized under the release protocol. This coverage is a property of source access, metadata matching, and parsing, not a model-performance metric. Table 6 records the 74-instance unified oracle-evaluable subset, the stable materialized subset, and the run-specific materialization differences. The main oracle table reports the completed oracle runs and should therefore be read as an upper-bound-style diagnostic rather than a perfectly controlled same-subset rerun.

Table 7. **Planning-window and assisted-query ablations.** Values are percentages with counts in parentheses. Raw planning is measured under different context renderings; assisted query accuracy uses the sanitized anchor and measures whether the generated query targets the intended tool family.

	Raw planning: $\mathrm{AnchorAcc}$				Assisted query
Model	local only $m = 5$	global+local $m = 1$	global+local $m = 3$	global+local $m = 5$	$\mathrm{QueryAcc}$ $m = 5$
GPT-5.2	10.5% (21/200)	8.0% (16/200)	6.0% (12/200)	5.5% (11/200)	18.5% (37/200)
Gemini 3.1 Pro	6.5% (13/200)	4.0% (8/200)	3.5% (7/200)	2.0% (4/200)	11.5% (23/200)
Claude Opus 4.5	23.0% (46/200)	30.0% (60/200)	23.5% (47/200)	24.0% (48/200)	41.0% (82/200)
DeepSeek V3.2	11.5% (23/200)	7.0% (14/200)	10.5% (21/200)	8.0% (16/200)	14.0% (28/200)
Qwen3-235B Thinking	41.5% (83/200)	41.5% (83/200)	33.0% (66/200)	40.5% (81/200)	30.0% (60/200)
Kimi K2 Thinking	44.5% (89/200)	44.5% (89/200)	**52.0%** (104/200)	43.5% (87/200)	48.0% (96/200)
Grok 4	24.4% (48/197)	27.0% (54/200)	27.5% (55/200)	26.5% (53/200)	**55.6%** (110/198)

Note. Denominators differ for two Grok diagnostic cells because a small number of diagnostic judgments were unavailable.

D. Additional Results and Diagnostics

This section contains diagnostics that support the main results without changing the leaderboard definition: planning and query ablations, uncertainty estimates, domain stratification, single-annotator audit results, and source-level failure decomposition. Completed metric tables use the seven model runs reported in the main text. The audit results below are human-confirmed single-annotator checks; they are not an inter-annotator agreement study.

D.1. Planning-Window and Query Ablations

Planning diagnostics evaluate whether a Raw-track model can infer a sanitized proof intention before retrieval. Query diagnostics evaluate whether an Assisted-track query targets the intended tool family after the anchor is supplied. These metrics localize failures before source selection and extraction; they are not part of the primary leaderboard.

Interpretation. The planning ablation is context-sensitive but not monotone. GPT-5.2 and Gemini score higher with local-only context than with global+local $m = 5$, suggesting that global setup can distract from the immediate proof transition. Claude peaks at global+local $m = 1$, Kimi peaks at $m = 3$, Qwen remains high across local-only and $m = 5$, and Grok is comparatively stable. The assisted-query diagnostic exposes a later bottleneck: Grok, Kimi, Claude, and Qwen reach 55.6%, 48.0%, 41.0%, and 30.0% QueryAcc, but their end-to-end $\mathrm{ToolAcc}$ values remain 3.5%, 3.5%, 7.0%, and 1.0%. Thus an on-target search intention is not sufficient for selecting, reading, and applying the right mathematical tool. These diagnostics explain upstream errors; they do not replace $\mathrm{ToolAcc}$, which scores the final source–tool pair. Because the construction audit identifies residual anchor-sanitization failures, Assisted $\mathrm{QueryAcc}$ should be interpreted as a diagnostic under imperfect anchors rather than as a leaderboard metric. The Raw end-to-end $\mathrm{ToolAcc}$ condition never exposes anchor text, and therefore is not directly affected by anchor leakage.

D.2. Statistical Uncertainty

Because Eval-200 contains multiple gaps from some citing papers, we report both ordinary binomial intervals and paper-cluster bootstrap intervals. Table 8 gives uncertainty estimates for $\mathrm{ToolAcc}$. Full metric-level intervals are included in the aggregation artifact.

D.3. Domain Stratification

Eval-200 is balanced by domain, with 40 instances in each of five broad areas. Table 9 reports $\mathrm{ToolAcc}$ by domain. Because each cell contains only 40 instances, the table should be read as a diagnostic view rather than a stable domain ranking.

Table 8. **Uncertainty for** ToolAcc**.** Wilson intervals use instance-level binomial uncertainty; paper-cluster intervals resample citing papers.

Model	ToolAcc	Wilson 95% CI	Paper-cluster 95% CI
Claude Opus 4.5	7.0% (14/200)	[4.2%, 11.4%]	[3.6%, 10.7%]
Grok 4	3.5% (7/200)	[1.7%, 7.0%]	[1.0%, 6.6%]
Kimi K2 Thinking	3.5% (7/200)	[1.7%, 7.0%]	[1.0%, 6.4%]
GPT-5.2	3.0% (6/200)	[1.4%, 6.4%]	[1.0%, 5.5%]
DeepSeek V3.2	2.5% (5/200)	[1.1%, 5.7%]	[0.5%, 5.0%]
Gemini 3.1 Pro	2.0% (4/200)	[0.8%, 5.0%]	[0.5%, 4.0%]
Qwen3-235B Thinking	1.0% (2/200)	[0.3%, 3.6%]	[0.0%, 2.6%]

Table 9. **Domain-level** ToolAcc **on Eval-200.** Each domain has 40 instances.

Model	**Alg./NT**	**Anal./PDE**	**Comb.**	**Geom./Top.**	**Prob./Stat./Ctrl.**
GPT-5.2	0.0%	5.0%	5.0%	2.5%	2.5%
Gemini 3.1 Pro	0.0%	5.0%	0.0%	2.5%	2.5%
Claude Opus 4.5	5.0%	2.5%	10.0%	7.5%	10.0%
DeepSeek V3.2	0.0%	0.0%	5.0%	0.0%	7.5%
Qwen3-235B Thinking	0.0%	0.0%	0.0%	2.5%	2.5%
Kimi K2 Thinking	0.0%	0.0%	7.5%	2.5%	7.5%
Grok 4	0.0%	5.0%	5.0%	2.5%	5.0%

D.4. Single-Annotator Judge Audit and Source-Level Failure Decomposition

The main metrics depend on the GROUND and SUFF predicates, so we audited a stratified sample of judge outputs. A single human annotator confirmed assistant-prepared signoff judgments on 2026-04-26. This audit is not an inter-annotator agreement study: the reported agreement and κ compare the fixed judge labels against the single human-confirmed audit labels.

Alternative-source successes are central to the citation-invariant claim, so the audit also checks their admissibility. The human-confirmed audit validates 39 of 42 judge-scored alternative-source successes and all 3 cited-source successes. The three invalid alternative-source successes are false positives for sufficiency or source admissibility; Table 13 records their impact as an audit sensitivity analysis. The main reported aggregate tables remain the fixed-judge benchmark scores, while this audit quantifies their residual label risk.

Two additional cases were protocol flags rather than predicate disagreements: `1909.11981v3_gap_3` for Claude and `2109.07887v3_gap_1` for Qwen selected sources matching the citing paper or its title. These flags motivate the explicit citing-paper exclusion rule in Appendix C.2 and should be revisited in any future audit-corrected scoring artifact.

The paper-facing failure decomposition is source-level rather than citation-level. Because the task is citation-invariant, failures are not defined by whether the author-cited source appears in Top-20. Table 14 partitions final outcomes into no grounded support, grounded but insufficient, and success. Retrieval-conditioned views based on CiteRecall@20 remain useful for debugging, but they are not the official failure definition.

Table 10. **Single-annotator audit of judge labels.** The audit covers 175 sampled outputs stratified by model and predicted outcome. Eligible labels exclude abstentions and cases without an evaluable prediction; here abstentions are recorded as `uncertain` by both the fixed judge and the human-confirmed audit label. κ is judge-vs-human-confirmed kappa, not inter-annotator kappa.

Predicate	Eligible	Agreement	Precision	Recall	F1	κ
GROUND	156	100.0%	100.0%	100.0%	100.0%	1.000
SUFF	156	98.1%	93.3%	100.0%	96.6%	0.952
TOOLACC	155	98.1%	93.3%	100.0%	96.6%	0.952

Table 11. **Judge-audit sample composition.** The 175 audited outputs are stratified across success, grounded-but-insufficient, ungrounded, and judge-abstained buckets as well as across the seven reported models.

Bucket	GPT	Gemini	Claude	DeepSeek	Qwen	Kimi	Grok	Total
Success	6	4	14	5	2	7	7	45
Grounded but insufficient	10	4	8	12	9	9	18	70
Ungrounded	5	7	6	5	8	5	4	40
Judge abstained	3	3	2	3	4	3	2	20
Total	24	18	30	25	23	24	31	175

E. Controlled Retrieval vs. Live Web-Browsing Agents

Recent "deep research" systems perform long-horizon web research by planning searches, issuing iterative queries, reading many pages or files, and synthesizing cited reports (OpenAI, 2025; Google AI for Developers, 2026). Academic benchmarks increasingly study such agents as report writers or general web researchers (FutureSearch et al., 2025; Li et al., 2025). We view these systems as complementary to Re^{2}Math, but outside the primary fixed-budget, fixed-backend leaderboard.

Rationale. Unrestricted browsing introduces variables that are orthogonal to the mathematical policy under evaluation: live-search drift, proprietary indices, ranking and crawling policy, page-opening strategy, context compression, retry budget, freshness filtering, and provider-specific safety layers. These hidden operations overlap with the benchmarked stages $\pi_{\mathrm{qry}}, \pi_{\mathrm{sel}}, \pi_{\mathrm{ext}}$. Allowing each model to use a different integrated search stack would therefore compare model-plus-backend systems rather than the controlled source–tool retrieval policy.

Benchmark contract. The primary protocol follows the information-retrieval practice of shared test collections and standardized evaluation conditions (Sanderson, 2010): agents see the same instances, candidate budget, cached query outputs, and verification policy. Models still decide what to search for and how to use retrieved evidence, but query execution, candidate ranking, source materialization, context budget, and logs are supplied by the release-fixed backend.

Agentic-learning view. The same interface can be viewed as a finite-horizon agentic decision process. The state contains the rendered proof context, the model's current planning description, previous queries, cached retrieval results, and, after a source-opening action, the fixed source materialization. The action space consists of planning, query formulation, metadata-based source selection, optional source opening, extraction or faithful restatement of a theorem-like statement, and abstention. The terminal reward is the source–tool success indicator

$$\mathbf{1}[\mathsf{Ground}_i(\hat{d}_i, \hat{K}_i) \wedge \mathsf{Suff}_i(\hat{K}_i \mid \mathcal{C}_i, \Delta_i)],$$

with auxiliary diagnostics such as $\mathrm{AnchorAcc}$, $\mathrm{CiteRecall@20}$, and $\mathrm{GroundRate}$ available as structured feedback. Because query results and source materializations are frozen, this environment supports reproducible offline policy analysis and future agentic RL studies without conflating policy learning with live-search drift or proprietary browsing stacks. The present paper uses this structure for evaluation only; it does not train policies on Re^{2}Math.

Table 12. **Human audit of successful source type.** The audit covers all judge-scored successful outputs in Eval-200.

Success type	Valid / audited	Validity
Alternative-source successes	39 / 42	92.9%
Cited-source successes	3 / 3	100.0%

Table 13. **Audit-flagged invalid alternative-source successes.** These cases are useful for sensitivity analysis because they show how much the fixed-judge ToolAcc would move if the single-annotator audit labels were applied retroactively.

Model	Instance	Audit reason	If corrected
GPT-5.2	`2012.12831_gap_0`	Recovered the Boolean-to-tropical transfer theorem, but not the monotone lower bound needed to instantiate it.	ToolAcc -1; AltSource-ToolAcc -1
GPT-5.2	`2401.14380v1_gap_0`	Retrieved the paper's own target theorem rather than the external spline-formula witness used for the proof step.	ToolAcc -1; AltSource-ToolAcc -1
Claude Opus 4.5	`2010.01779v3_gap_0`	Retrieved only the heavy-tail trap assumption; the audited witness is a scaling-limit theorem.	ToolAcc -1; AltSource-ToolAcc -1

Citation-derived shortcut control. Because instances are mined from explicit proof citations, unrestricted browsing can reward bibliographic shortcuts such as phrase matching, citation-chain following, or direct recovery of the author-cited paper. Such behavior may be useful in deployed research assistants, but it changes the benchmark target from identifying a sufficient mathematical tool to open-web source hunting.

Future browsing track. A future controlled browsing or native-search track should use the same frozen release snapshot, or a separately released browsing snapshot, and should report query budgets, visited pages, opened files, backend details, context-construction rules, and complete interaction logs alongside TOOLACC.

E.1. Implementation Notes

Several helper-script options are not part of the official benchmark definition. Solver-as-judge fallback is only a debugging convenience; all paper-facing runs use the fixed held-out judge. Legacy proxy fields such as `tool_success_proxy`, `is_matched`, and `cited_tool_success_proxy` are not paper metrics. The official retrieval date policy is the Google Scholar year-level cutoff specified in Appendix C.1.

F. Detailed Case Studies

This appendix gives a mechanism-level reading of representative instances from the completed seven-model evaluation. The cases are not an additional split and should not be read as anecdotal successes or failures; they are qualitative probes of the aggregate pattern in Section 6. Each case first presents the proof situation and reference witness, then reports the observed model behavior, and finally explains which interface breaks: proof need $\rightarrow$ search-effective language, retrieved source $\rightarrow$ theorem-bearing passage, or sourced statement $\rightarrow$ assumption-checked proof tool.

Legend. Y/N indicate positive/negative labels, and $\perp$ indicates no evaluable output. **Alt** means that a successful source was not the stored author-cited source. **Plan** is the five-block Raw-track planning match. **Query** is the assisted-query diagnostic. Titles are shortened for readability.

F.1. Robust Alternative-Source Success: Brownian Random Interlacements

Instance. `1804.05967v2_gap_0`. **Paper.** *Two-dimensional Brownian random interlacements*. **Domain.** Probability / statistics / control.

Proof situation. The proof uses a comparison between planar Brownian motion conditioned to hit a larger circle before a smaller one and a Doob h-transformed process.

Reference witness. For all $R > 1$ and $1 < \|x\| < R$,

$$\mathbb{P}_x\big[W|_{[0,\tau(R)]} \in \cdot \mid \tau(R) < \tau(1)\big] = \mathbb{P}_x\big[\widehat{W}|_{[0,\widehat{\tau}(R)]} \in \cdot\big].$$

Construction citation. Comets–Popov–Vachkovskaia, *Two-dimensional random interlacements and late points for random walks*.

Table 14. **Source-level failure decomposition on Eval-200.** "Grounded but insufficient" equals $\mathrm{GroundRate} - \mathrm{ToolAcc}$ under the fixed-judge benchmark scores.

Model	No grounded support	Grounded but insufficient	Success
GPT-5.2	150/200 (75.0%)	44/200 (22.0%)	6/200 (3.0%)
Gemini 3.1 Pro	176/200 (88.0%)	20/200 (10.0%)	4/200 (2.0%)
Claude Opus 4.5	152/200 (76.0%)	34/200 (17.0%)	14/200 (7.0%)
DeepSeek V3.2	142/200 (71.0%)	53/200 (26.5%)	5/200 (2.5%)
Qwen3-235B Thinking	166/200 (83.0%)	32/200 (16.0%)	2/200 (1.0%)
Kimi K2 Thinking	152/200 (76.0%)	41/200 (20.5%)	7/200 (3.5%)
Grok 4	117/200 (58.5%)	76/200 (38.0%)	7/200 (3.5%)

Model behavior.

Model	Query	Plan	Query	Cite@20	Ground	Suff	Tool	Selected source
GPT-5.2	Brownian sausage covering	N	N	N	N	N	N	LIL for cover times by Wiener sausage
Gemini 3.1 Pro	boundedness vacant set Brownian interlacements	N	N	N	N	N	N	Two-Dimensional Brownian Random Interlacements
Claude Opus 4.5	random interlacement covering probability asymptotics	N	N	N	Y	Y	Y	Two-Dimensional Brownian Random Interlacements
DeepSeek V3.2	BRI sausages covering probability limit	N	N	N	Y	Y	Y	Two-Dimensional Brownian Random Interlacements
Qwen3-235B Thinking	BRI sausages compact covering probability	Y	N	N	Y	Y	Y	Two-Dimensional Brownian Random Interlacements
Kimi K2 Thinking	h-transform transition kernel Brownian	Y	N	N	Y	Y	Y	Two-dimensional Brownian random interlacements
Grok 4	BRI covering probability alpha infinity	N	N	N	Y	Y	Y	Two-Dimensional Brownian Random Interlacements

Representative extraction. Claude, DeepSeek, Qwen, Kimi, and Grok all recover a statement equivalent to Lemma 2.1 of the selected source: Brownian motion stopped at $\tau(R)$, conditioned on $\tau(R) < \tau(1)$, has the same law as the h-transformed process stopped at $\widehat{\tau}(R)$. The notation varies, but the assumptions and conclusion match the reference witness.

Mechanistic diagnosis. This case isolates why citation-invariance is a mathematical requirement rather than an evaluation convenience. All successful systems have $\mathrm{CiteRecall@20} = 0$, yet their selected source states precisely the path-law identity needed by the proof. Exact citation recovery would therefore penalize valid scholarly justifications. At the same time, the upstream planning and query signals are mostly negative, showing that the final object must be judged as a source–tool pair rather than as a sequence of intermediate proxy labels.

Takeaway. The lesson is that provenance and applicability are separable from citation identity. The author citation remains useful for construction and debugging, but the success criterion must ask whether the returned source supports a sufficient tool. This is exactly the distinction between $\mathrm{CiteRecall@20}$ as a diagnostic and $\mathrm{ToolAcc}$ as the task metric.

F.2. Citation-Hit Classical Theorem Success: Positive Definite Functions on Spheres

Instance. `1901.06096v3_gap_1`. **Paper.** *Repeated minimizers of p-frame energies*. **Domain.** Combinatorics / discrete mathematics.

Proof situation. The proof reduces a p-frame energy bound to the case $p = 2$, where it invokes positive definite functions on the unit sphere.

Reference witness. A continuous function $f : [-1, 1] \to \mathbb{R}$ is positive definite on $\mathbb{S}^{d-1}$ iff it admits a Gegenbauer expansion $f(t) = \sum_{k \geq 0} a_k C_k^{(d-2)/2}(t)$ with $a_k \geq 0$.

Construction citation. Schoenberg, *Positive definite functions on spheres*, 1941.

Model behavior.

Table 15. **Case-study roadmap.** Each case isolates a distinct mechanism behind low end-to-end ToolAcc.

Case	Mechanism	Reference tool	Diagnostic readout
1804.05967v2 _gap_0	Robust alternative-source success	Conditioned planar Brownian motion equals an h-transform	Mathematical validity is not bibliographic identity: several systems close the gap while missing the author citation.
1901.06096v3 _gap_1	Citation-hit classical theorem success	Schoenberg characterization of positive definite functions on spheres	Citation recall can be high while theorem-level localization and formulation remain fragile.
1510.08891 _gap_1	Oracle rescue / retrieval bottleneck	Matroidal locus equals the Voronoi–perfect-cone intersection	The proof vocabulary points toward cubic threefolds, but the needed tool lives in toroidal-fan language.
1804.10581v2 _gap_12	Stronger theorem succeeds; nearby theorem fails	Non-controllability of fractional heat equation	A correct topical source is not enough; the equation, domain, parameter range, and quantifiers must match.
1909.12545 _gap_1	Grounded but insufficient local corollary	Singular $\mathbb{Q}$-factorial terminal varieties admit no proper crepant resolution	Source-supported corollaries fail when their extra hypotheses are not established by the proof state.
2507.02263v2 _gap_1	Weaker theorem under same assumptions	Signless-Laplacian excess gives $\Omega(n^{k+1})$ cliques	The same assumptions can support multiple conclusions; only the right strength closes the gap.
1905.05370 _gap_0	Anchor consensus, query collapse	Compactness/convergence theorem for Muskat approximations	Models recognize compactness but compress the proof obligation into overly generic method words.
2012.12831 _gap_0	Semantically plausible query, retrieval/protocol ambiguity	Monotone lower bound for the logical permanent	Internal transfer principles can look useful while missing the external theorem that instantiates them.
2110.11087v2 _gap_1	Canonical algebra query, zero retrieval	Polynomial gluing for Steinberg groups	Even nearly canonical theorem-family language may not surface an older, poorly indexed algebra source.

Model	Query	Plan	Query	Cite@20	Ground	Suff	Tool	Selected source
GPT-5.2	positive definite polynomial sphere	N	Y	Y	$\perp$	$\perp$	$\perp$	Strictly positive definite functions on spheres
Gemini 3.1 Pro	positive definite function coefficients one	N	Y	Y	N	N	N	Metric spaces and positive definite functions
Claude Opus 4.5	positive definite functions spheres Gegenbauer	Y	Y	Y	Y	Y	Y	Positive definite functions in distance geometry
DeepSeek V3.2	positive definite functions unit sphere	Y	Y	Y	$\perp$	$\perp$	$\perp$	none
Qwen3-235B Thinking	spherical positive definite sum squares	Y	N	Y	Y	N	N	Positive definite functions on the unit sphere and integrals of Jacobi polynomials
Kimi K2 Thinking	positive definite sphere quadratic forms	N	Y	Y	N	N	N	Optimal measures for p-frame energies
Grok 4	positive definite functions Gegenbauer expansion	Y	Y	Y	N	N	N	Optimal measures for p-frame energies

Representative success. Claude extracts a Schoenberg-type theorem: a continuous real function on the sphere is positive definite iff it is a nonnegative linear combination of Gegenbauer polynomials. The source uses slightly different notation, but the statement is mathematically equivalent to the witness and directly supplies the positive-definiteness tool for the $p = 2$ energy argument.

Representative near miss. Qwen selects a related paper and grounds a true sufficient-condition theorem for positive definiteness involving smoothness and derivative sign conditions. The theorem is real, but it is not the Schoenberg equivalence and it imposes hypotheses not established in the proof context. It therefore fails sufficiency.

Mechanistic diagnosis. This is a relatively favorable retrieval case: the intended theorem family is classical, most queries contain the right mathematical vocabulary, and the stored citation often appears in the candidate pool. The difficulty therefore shifts downstream. The model must recognize that the proof needs Schoenberg's equivalence theorem, not merely any

positive-definiteness criterion or a paper about spherical kernels. Only Claude turns the retrieved literature into a statement with the right iff structure, polynomial basis, and nonnegative-coefficient condition.

Takeaway. This case shows that source acquisition is only a coarse gate. A theorem-use agent must operate below the paper level: it must identify the theorem boundary, preserve the logical form, and decide whether the extracted statement has the right level of generality for the local argument.

F.3. Oracle Rescue: Matroidal Locus and Cubic Threefold Compactifications

Instance. `1510.08891_gap_1`. **Paper.** *Complete moduli of cubic threefolds and their intermediate Jacobians*. **Domain.** Algebra / number theory.
Proof situation. The proof needs a toroidal-compactification fact to show that the extended intermediate-Jacobian map lands in the correct compactification locus.
Reference witness. The matroidal locus $\mathrm{Matr}[5]$ is the maximal partial compactification contained in both the second Voronoi and perfect-cone compactifications; equivalently, $\Sigma^V \cap \Sigma^P = \Sigma^{\mathrm{mat}}$.
Construction citation. Melo–Viviani, *Comparing perfect and 2nd Voronoi decompositions: the matroidal locus*.

Model behavior.

Model	Query	Plan	Query	Cite@20	Ground	Suff	Tool	Selected source
GPT-5.2	cubic threefold monodromy cones	N	N	N	N	N	N	Moduli space via degenerations
Gemini 3.1 Pro	local monodromy boundary divisors	N	N	N	N	N	N	A p-adic local monodromy theorem
Claude Opus 4.5	monodromy intermediate Jacobian cubic threefold	N	N	N	$\perp$	$\perp$	$\perp$	none
DeepSeek V3.2	monodromy intermediate Jacobian cubic threefolds	N	N	N	N	N	N	Moduli space via degenerations
Qwen3-235B Thinking	monodromy GIT boundary Voronoi compatibility	N	N	N	$\perp$	$\perp$	$\perp$	none
Kimi K2 Thinking	monodromy intermediate Jacobian GIT boundary	N	N	N	N	N	N	Moduli space via degenerations
Grok 4	monodromy intermediate Jacobian cubic threefolds	N	N	N	Y	N	N	Moduli space via degenerations

Representative near miss. Grok grounds a statement about the boundary geometry of intermediate Jacobians. The statement is related to the citing paper, but it is not the fan-theoretic equality $\Sigma^V \cap \Sigma^P = \Sigma^{\mathrm{mat}}$ and does not imply the maximality of the matroidal partial compactification.

Oracle rescue. In the oracle-source condition, GPT and Claude recover the intended statement from Melo–Viviani. Claude extracts the clean fan identity: a cone belongs to both the second Voronoi and perfect-cone decompositions iff it belongs to the matroidal decomposition.

Mechanistic diagnosis. Open retrieval is captured by the surface vocabulary of the citing proof: cubic threefolds, monodromy, degenerations, and intermediate Jacobians. The external tool, however, is expressed in a different mathematical register: Voronoi and perfect-cone decompositions, fan intersections, and the matroidal locus. Oracle access changes the problem qualitatively: once Melo–Viviani is supplied, GPT and Claude can recover the relevant fan identity. The failure is therefore not that the tool is unreadable, but that the model cannot bridge from the local proof vocabulary to the external theorem family.

Takeaway. The case demonstrates a central source-acquisition problem: the best search terms for a proof step may not be the nouns appearing in the local proof. The model must infer the hidden mathematical abstraction that the authors import from another literature.

F.4. Same Topic, Wrong Theorem: Fractional Heat Non-Controllability

Instance. `1804.10581v2_gap_12`. **Paper.** *Lack of null-controllability for the fractional heat equation and related equations*. **Domain.** Probability / statistics / control.

Proof situation. The proof needs non-controllability for the fractional heat equation on $\Omega = \mathbb{R}$ or $\mathbb{T}$, for $0 \leq \alpha < 1$, on any strict open control subset.

Reference witness. If $0 \leq \alpha < 1$, $\Omega = \mathbb{R}$ or $\mathbb{T}$, and $\omega \subsetneq \Omega$ is open, then the fractional heat equation is not null-controllable on ω in any time $T > 0$.

Construction citation. The citing paper points to an earlier special case for a related phase-space domain and time threshold.

Model behavior.

Model	Query	Plan	Query	Cite@20	Ground	Suff	Tool	Selected source
GPT-5.2	fractional heat controllability counterexamples	N	N	N	Y	N	N	Lack of null-controllability...
Gemini 3.1 Pro	fractional heat lack null controllability	N	N	N	Y	N	N	Lack of null-controllability...
Claude Opus 4.5	controllability cost blow-up time threshold	N	N	N	$\perp$	$\perp$	$\perp$	none
DeepSeek V3.2	fractional heat null controllability decay	N	N	N	Y	Y	Y	Lack of null-controllability...
Qwen3-235B Thinking	high-frequency non-observability fractional heat equation	Y	N	N	N	N	N	Seḿinaire de théorie spectrale...
Kimi K2 Thinking	fractional heat observability inequality concentration	N	N	N	$\perp$	N	$\perp$	Exact observability properties...
Grok 4	fractional heat equation null-controllability minimal	N	N	N	Y	N	N	Lack of null-controllability...

Successful extraction. DeepSeek extracts Theorem 1.3 from the selected source: for $0 \leq \alpha < 1$, $\Omega = \mathbb{R}$ or $\mathbb{T}$, and strict open ω, the fractional heat equation is not null-controllable in any time. This is stronger than the special-case witness and therefore sufficient.

Near misses. GPT-5.2, Gemini, and Grok select a topical source and ground real non-controllability or observability statements, but the statements concern a different equation, phase-space domain, or time threshold. Qwen's query captures high-frequency non-observability but selects an unrelated source.

Mechanistic diagnosis. This case makes the sufficiency predicate concrete. The retrieved papers are not random distractors; they contain genuine observability and non-controllability statements in the same topical neighborhood. What separates success from failure is exact applicability: the equation class, state space, control set, parameter regime, and quantification over time must all line up with the local proof state. DeepSeek succeeds because it extracts a statement whose scope covers the target proof obligation, whereas several grounded near misses fail on one of these dimensions.

Takeaway. This is the pattern behind the large $\mathrm{GroundRate}$-to-$\mathrm{ToolAcc}$ gap in the main tables. Grounding answers the provenance question, but the proof requires a second judgment: whether the sourced statement is the right mathematical instrument for this exact transition.

F.5. Grounded but Insufficient: Symplectic Resolutions and Birational Obstructions

Instance. `1909.12545_gap_1`. **Paper.** *Symplectic resolutions of character varieties*. **Domain.** Geometry / topology.

Proof situation. The proof wants to rule out proper symplectic resolutions of singular character varieties.

Reference witness. A singular $\mathbb{Q}$-factorial terminal variety admits no proper crepant resolution; hence, in the symplectic setting, no proper symplectic resolution.

Why this case matters. It separates a general birational obstruction from local corollaries that require extra representation-theoretic hypotheses.

Model behavior.

Model	Query	Plan	Query	Cite@20	Ground	Suff	Tool	Selected source
GPT-5.2	Q-factorial terminal crepant resolution	N	N	N	Y	N	N	Symplectic resolutions and factoriality...
Gemini 3.1 Pro	Q-factorial terminal crepant resolution	N	Y	N	N	N	N	Global moduli theory of symplectic varieties
Claude Opus 4.5	terminal singularities crepant resolution obstruction	N	Y	N	Y	Y	Y	Symplectic resolutions of character varieties
DeepSeek V3.2	character variety singular locus codimension	N	N	N	Y	N	N	Symplectic resolutions of character varieties
Qwen3-235B Thinking	character variety singular locus codimension	N	N	N	Y	N	N	On the codimension of the singular locus
Kimi K2 Thinking	character variety singular locus Q-factorial	Y	Y	N	Y	N	N	Symplectic resolutions of character varieties
Grok 4	character variety singular locus codimension	Y	Y	N	N	N	N	On the codimension of the singular locus

Successful extraction. Claude finds a statement equivalent to the general obstruction: a singular $\mathbb{Q}$-factorial terminal variety does not admit a proper crepant resolution, and therefore no symplectic resolution in the relevant setting.

Near misses. GPT-5.2, DeepSeek, Qwen, and Kimi ground statements in related sources, but the statements are either specialized to Hamiltonian reductions, codimension estimates, or local facts whose extra hypotheses are not established in the proof context.

Mechanistic diagnosis. The proof needs a general birational obstruction: terminal $\mathbb{Q}$-factorial singularities rule out proper crepant, hence symplectic, resolutions. Several systems instead extract narrower corollaries from the character-variety literature. These statements are source-supported and locally plausible, but their hypotheses are not supplied by the proof state. The error is therefore a mismatch in logical portability: the model finds a theorem in the right neighborhood, but not one that can be transported into the current argument.

Takeaway. The case shows why sufficiency must be judged relative to the local proof context. A theorem can be real, relevant, and still unusable if its assumptions encode a more specialized situation than the citing proof has established.

F.6. Weaker Theorem Under Similar Assumptions: Signless-Laplacian Supersaturation

Instance. `2507.02263v2_gap_1`. **Paper.** *Some Turán-type results for the signless Laplacian spectral radius*. **Domain.** Combinatorics / discrete mathematics.
Proof situation. The proof requires a quantitative clique-counting theorem under a signless-Laplacian spectral-radius assumption.
Reference witness. If $k \geq 2$, $\varepsilon > 0$, and $q(G) \geq (1 - 1/k + \varepsilon)2n$, then G contains $\Omega_{k,\varepsilon}(n^{k+1})$ copies of K_{k+1}.
Why this case matters. It distinguishes a theorem that produces a large blow-up from the stronger counting statement actually needed by the proof.

Model behavior.

Model	Query	Plan	Query	Cite@20	Ground	Suff	Tool	Selected source
GPT-5.2	signless Laplacian supersaturation blow-up	Y	N	N	Y	N	N	Signless Laplacian spectral radius
Gemini 3.1 Pro	blow-up graph subgraph containment	N	N	N	N	N	N	Graph containment problems
Claude Opus 4.5	spectral supersaturation complete multipartite subgraph	Y	N	N	Y	Y	Y	Signless Laplacian spectral radius
DeepSeek V3.2	signless Laplacian spectral radius complete	N	N	N	$\perp$	$\perp$	$\perp$	none
Qwen3-235B Thinking	complete $k+1$-partite linear-sized parts	N	N	N	N	N	N	Linear-sized minors...
Kimi K2 Thinking	signless Laplacian supersaturation chromatic number	Y	Y	N	Y	Y	Y	Signless Laplacian spectral radius
Grok 4	blow-up embeddings subgraph counting chromatic	Y	N	N	N	N	N	Counting Graph Homomorphisms

Successful extraction. Claude and Kimi extract the quantitative counting theorem: spectral excess above the Turán threshold forces $\Omega(n^{k+1})$ copies of K_{k+1}, which is exactly the scale required downstream.

Near miss. GPT-5.2 grounds a true theorem from the same paper family, but it only yields a suitable blow-up with logarithmic-size parts. That theorem can imply existence of certain subgraphs but not the polynomial number of clique copies needed by the proof.

Mechanistic diagnosis. Here the difficulty is not hypothesis mismatch but conclusion strength. The successful statements and the near miss inhabit the same signless-Laplacian/Turán family, and the assumptions look similar. The proof, however,

needs polynomially many K_{k+1} copies. A blow-up or qualitative containment theorem is too weak: it can show existence-like structure without supplying the counting scale used downstream.

Takeaway. Assumption checking is not only about preconditions. It also includes the force of the conclusion: asymptotic order, uniformity, constants, quantifiers, and whether a statement is qualitative or quantitative. This is a theorem-use skill that ordinary relevance retrieval does not measure.

F.7. Anchor Consensus, Query Collapse: Muskat Compactness

Instance. `1905.05370_gap_0`. **Paper.** *Weak Solutions to the Muskat Problem with Surface Tension Via Optimal Transport*. **Domain.** Analysis / PDE.

Proof situation. The proof needs a compactness theorem upgrading uniform bounds and time-translation control into strong L^1 convergence of approximate densities.

Reference witness. A convergence theorem for thresholding/variational approximations that gives the strong compactness required to pass to the limit.

Why this case matters. It shows that planning can be correct while query generation collapses to generic method words.

Model behavior.

Model	Query	Plan	Query	Cite@20	Ground	Suff	Tool	Selected source
GPT-5.2	Aubin Lions Wasserstein compactness	Y	N	N	$\perp$	N	$\perp$	Nonlinear evolution systems...
Gemini 3.1 Pro	Aubin-Lions compactness L^1 convergence	Y	N	N	Y	N	N	Aubin-Lions-Dubinskiĭ lemmas
Claude Opus 4.5	Aubin-Lions compactness Wasserstein gradient flow	Y	N	N	Y	N	N	Fourth-order PDE weak solutions
DeepSeek V3.2	strong L^1 convergence compactness Aubin-Lions	Y	N	N	Y	N	N	Aubin-Lions-Dubinskiĭ lemmas
Qwen3-235B Thinking	adapted compactness density strong L	Y	N	N	Y	N	N	Compactness in adapted weak topologies
Kimi K2 Thinking	Aubin-Lions compactness Wasserstein Muskat densities	Y	Y	N	Y	N	N	Fourth-order PDE weak solutions
Grok 4	Aubin-Lions compactness strong L^1 convergence	Y	N	N	Y	N	N	Aubin-Lions-Dubinskiĭ lemmas

Mechanistic diagnosis. All seven models recognize the broad proof intention, so this is not a planning failure in the usual sense. The failure occurs when the proof obligation is compressed into search language. Most systems translate the need into generic "Aubin–Lions" or compactness terminology, which retrieves real compactness sources and sometimes grounded statements, but not the theorem used by the Muskat proof. The lost information is discriminative: the approximation scheme, topology, equation class, and convergence mode that separate the cited tool from a large family of method-level compactness results.

Takeaway. The case explains why $\mathrm{AnchorAcc}$ and query diagnostics cannot substitute for $\mathrm{ToolAcc}$. Models can name the right kind of method while failing to preserve the features that make a search query theorem-bearing rather than merely topical.

F.8. Semantically Plausible Query, Protocol-Sensitive Success: Boolean-to-Tropical Lower Bounds

Instance. `2012.12831_gap_0`. **Paper.** *Approximation Limitations of Pure Dynamic Programming*. **Domain.** Probability / statistics / control.

Proof situation. A Boolean lower bound for the assignment/permanent predicate must be instantiated through a Boolean-to-tropical transfer principle.

Reference witness. The logical permanent requires monotone Boolean circuits of size at least $n^{\Omega(\log n)}$.

Construction citation. Razborov, *Lower bounds on monotone complexity of the logical permanent*, 1985.

Model behavior.

Model	Query	Plan	Query	Cite@20	Ground	Suff	Tool	Selected source
GPT-5.2	tropical circuit boolean reduction	N	N	N	Y	Y	Y	Approximation Limitations of Pure Dynamic Programming
Gemini 3.1 Pro	Jerrum Snir monotone perfect matching	N	N	N	⊥	⊥	⊥	none
Claude Opus 4.5	tropical circuit monotone boolean simulation	Y	Y	N	N	N	N	Reciprocal inputs in arithmetic and tropical circuits
DeepSeek V3.2	boolean assignment problem lower bound	Y	Y	N	⊥	⊥	⊥	none
Qwen3-235B Thinking	tropical circuit monotone boolean approximation	Y	N	N	Y	N	N	Monotone circuit complexity of Boolean functions
Kimi K2 Thinking	Jerrum Snir permanent monotone circuit	Y	Y	N	⊥	⊥	⊥	none
Grok 4	monotone Boolean circuit lower bound	Y	Y	N	N	N	N	Monotone circuit complexity of Boolean functions

Audit-sensitive outcome. The fixed judge scores GPT-5.2 as successful, but the single-annotator audit flags this as an alternative-source false positive: GPT recovers the Boolean-to-tropical transfer theorem from the citing paper, not the monotone lower bound for the logical permanent needed to instantiate the transfer. This is one of the sensitivity cases reported in Appendix D.

Mechanistic diagnosis. This case exposes a boundary condition for source admissibility. The proof contains an internal Boolean-to-tropical transfer principle, but the citation supplies the monotone Boolean lower bound needed to instantiate that principle. GPT's fixed-judge success retrieves the internal transfer theorem from the citing paper, which is real and proof-relevant but not the external witness. The audit therefore flags it as a false positive under the external-source protocol. At the same time, the queries are semantically reasonable, indicating that the failure is not a trivial misunderstanding of the proof architecture.

Takeaway. The lesson is that source-grounded evaluation needs an explicit boundary between internal proof machinery and imported tools. Without that boundary, a model could receive credit for restating the citing paper's own bridge lemma while missing the external theorem that actually supplies the missing premise.

F.9. Canonical Algebra Query, Zero Retrieval: Tulenbaev Lifting

Instance. `2110.11087v2_gap_1`. **Paper.** *On the $\mathbb{A}^1$-invariance of K_2 modeled on linear and even orthogonal groups*.
Domain. Algebra / number theory.
Proof situation. The proof verifies axioms CFC–HIF for $K_2(\Phi, -)$. CFC and DP have been handled; the remaining step uses a Tulenbaev lifting/gluing property for Steinberg groups.
Reference witness. For a regular ring $R \supset k$ and $\Phi = A_\ell$, $\ell \geq 4$, the Steinberg group $\mathrm{St}(\Phi, R[t])$ satisfies the polynomial gluing property.
Construction citation. Tulenbaev, *The Steinberg group of a polynomial ring*, 1983.

Model behavior.

Model	Query	Plan	Query	Cite@20	Ground	Suff	Tool	Selected source
GPT-5.2	Tulenbaev lifting Steinberg	N	N	N	⊥	⊥	⊥	none
Gemini 3.1 Pro	Steinberg group Tulenbaev lifting property	N	N	N	⊥	⊥	⊥	none
Claude Opus 4.5	Steinberg group Tulenbaev lifting property	Y	Y	N	⊥	⊥	⊥	none
DeepSeek V3.2	K-homotopy invariance root system	N	N	N	⊥	⊥	⊥	none
Qwen3-235B Thinking	Steinberg groups Tulenbaev lifting simply-laced	N	Y	N	⊥	⊥	⊥	none
Kimi K2 Thinking	Tulenbaev lifting property Steinberg groups	N	Y	N	⊥	⊥	⊥	none
Grok 4	Steinberg groups Tulenbaev lifting property	Y	Y	N	⊥	⊥	⊥	none

Mechanistic diagnosis. This is the cleanest source-acquisition bottleneck in the case set. Several systems generate queries containing the distinctive author/tool vocabulary, including Tulenbaev, Steinberg groups, and lifting or gluing properties. Nevertheless, the shared backend does not surface the cited algebra source in the evaluated prefix, and no alternative source is selected. The model can be mathematically close at the planning and query levels while still failing because the relevant literature is old, terminology varies, or the theorem is poorly indexed.

Takeaway. The case motivates theorem-aware retrieval rather than simply larger language models. A better backend would index statements, assumptions, author/title variants, and theorem-level lexicalizations, so that canonical proof-language queries can reach the relevant source under a reproducible artifact contract.

F.10. Synthesis

Taken together, the cases sharpen the aggregate finding into three distinct failure interfaces. First, proof obligations must be lexicalized into terms that a scholarly index can surface; the Muskat and Tulenbaev cases show that even good mathematical intent can fail at this interface. Second, candidate documents must be converted into theorem-bearing passages; the Schoenberg and oracle-rescue cases show that citation recall or source access alone does not solve theorem localization. Third, sourced statements must be checked against the proof state; the fractional-heat, symplectic-resolution, and signless-Laplacian cases show that true statements can be unusable because of mismatched domains, assumptions, or conclusion strength. The Brownian case adds the complementary lesson that valid mathematical support need not coincide with the stored citation, while the Boolean-to-tropical case explains why the citing paper itself cannot be used as a shortcut for an external tool.

These examples support the aggregate conclusion in Section 6. The weakness is not a single missing module. Current LLMs can sometimes plan, sometimes retrieve, sometimes ground, and sometimes restate a theorem; what remains unreliable is the disciplined composition of those abilities into a proof-valid source–tool pair. This is the capability that $\mathrm{ToolAcc}$ measures and that Re^{2}Math is designed to make visible.

G. Limitations

Natural-language evaluation. Re^{2}Math evaluates source grounding and proof-gap sufficiency in natural-language mathematical documents. This is the right level for the research-assistance setting we target, but it is not a substitute for mechanically checked formal proof. A high TOOLACC score means that the returned statement is source-supported and judged sufficient for the local proof transition; it does not certify a formal derivation. The current validation includes a single-annotator judge-vs-human audit rather than a broad multi-annotator gold set or inter-annotator agreement study. Future releases should add adjudicated expert subsets, borderline-case annotations, and multiple admissible-source labels where feasible.

Retrieval artifact and source access. The primary end-to-end protocol uses a shared frozen scholarly-search artifact with Top-20 metadata inspection and at most one source opening. This design makes comparisons reproducible and exposes source-selection pressure, but it also means that failures can arise from model query choice, backend coverage, ranking behavior, source identity matching, or source materialization. We therefore interpret $\mathrm{CiteRecall@20}$ as an artifact-level diagnostic rather than as a pure measure of model mathematical ability or a complete classical IR evaluation. The same fixed interface also avoids comparing proprietary web-search, PDF parsing, OCR, equation rendering, or multimodal document-reading stacks in the main leaderboard. This improves auditability, but it is stricter than real research workflows and under-evaluates deployed agents that can inspect multiple sources, follow citation chains, or read rendered PDFs. Future tracks should report sensitivity to N, multiple source openings, theorem-aware retrieval backends, and standardized PDF-native or browsing budgets.

Benchmark construction and coverage. The current release is built from published, citation-observable proof gaps using a published-first, arXiv-backed pipeline. This favors proof-rich papers with recoverable source text and may underrepresent areas where key results appear primarily in books, older papers, surveys, folklore, publisher-only PDFs, or uncited background knowledge. Eval-200 is also balanced and quality-filtered rather than a random sample of all collected gaps, so the main numbers should be read as controlled diagnostic results on the held-out split rather than as population estimates over all mathematical research. The oracle-source diagnostic has a related boundary: $\mathrm{OracleCoverage}$ is a release-protocol materialization property, not a model score or a clean upper bound. Non-materialization can arise from ambiguous metadata, unavailable full text, book or survey sources, broken mirrors, or parser failures. The construction audit also shows residual citation-role instrumentality risk: in a minority of audited retained rows, the author citation is not a clean proof-tool witness, even though the audited reference-witness sufficiency and external-source checks pass. This limits the interpretation of the author citation as a historical witness, but not the definition of the official score, which is based on the final Ground and Suff predicates for the returned source–tool pair.

Scope of the task. Re^{2}Math intentionally isolates single-gap, source-grounded theorem/tool retrieval. It does not evaluate full autonomous research, theorem discovery, long-horizon proof repair, formalization, or unrestricted browsing. The stored witness is one sufficient route, not a complete characterization of all valid proof strategies; although citation-invariant scoring allows alternative sources, the sufficiency judge may still be conservative when a model closes the gap through a substantially

different mathematical route. Because the benchmark is built from published literature, contamination and memorization also cannot be fully ruled out, although source-grounded scoring reduces the value of unsupported memorized statements. The Assisted track is likewise diagnostic rather than primary, since residual anchor-sanitization risk remains; the Raw-track leaderboard does not expose anchors. These boundaries are deliberate: the benchmark narrows the problem enough to make provenance, source selection, and proof-gap sufficiency auditable, while leaving broader browsing, multimodal, and long-horizon agentic settings for future standardized tracks.