Scaffolding, Not Stronger Reasoning: Diagnosing Skill-Card Retrieval for Mathematical Problem Solving

He Ma [1] **Xinrui Chang** [1] **Yuhang Xiao** [2] **Ying Wang** [3]

Abstract

Skill-card retrieval is often presented as reusable reasoning memory: a solver retrieves compact procedural guidance distilled from prior solutions and conditions on it at test time. We ask whether this mechanism actually strengthens mathematical reasoning, or mainly acts as scaffolding for models that still need help. In a paired diagnostic study with released TRS-style DeepMath skill cards, retrieval improves weaker solvers substantially (+10.4 pp for GPT-4o-mini and +4.2 pp for GPT-5.4), but gives only a small and unreliable gain for GPT-5.5 (+0.6 pp). When exact source cards are removed, these gains do not persist, and the GPT-5.5 condition falls 2.6 pp below Direct. For the strongest solver in our study, lower-ranked and random cards shift paired flips toward harm, and manual audits show that many failures arise from partially applicable skills that introduce caveats, convention shifts, wrong limiting regimes, or symbolic drift. These findings suggest that skill-card retrieval is best viewed as useful scaffolding for weaker solvers rather than as a uniformly reliable improvement as solver capability increases.

1. Introduction

Mathematical reasoning has become a central testbed for modern language models, spanning natural-language problem solving, proof-assistant interaction, and inference-time verification or search (Hendrycks et al., 2021; Cobbe et al., 2021; Wei et al., 2022; Wang et al., 2023; Azerbayev et al., 2024; Yue et al., 2024; Yang et al., 2023; Xin et al., 2025; Tsoukalas et al., 2024). Yet small mistakes can invalidate plausible solutions, so AI-for-math systems need not only correct answers but also reliable decisions about when extra context is useful, redundant, or misleading. Retrieval-augmented solvers offer one form of test-time support: they condition on relevant examples, premises, proof states, or abstract strategies (Lewis et al., 2020; Karpukhin et al., 2020; Irving et al., 2016; Yang et al., 2023; Alshammari et al., 2026). Zhao et al. (2026) propose *Thinking with Reasoning Skills* (TRS), which compresses prior deliberation trajectories into compact *skill cards* that describe when a reasoning pattern applies, what steps to try, and what pitfalls to avoid. Prior results suggest that retrieved skills can improve accuracy while reducing completion length, making skill-card retrieval an appealing scaffolding mechanism for mathematical problem solving.

However, the value of retrieved skills is not fixed. As base reasoners become stronger, direct prompting already solves many problems that previously required additional scaffolding. In this regime, the question shifts from whether a retrieved skill can ever help to whether the gain should be interpreted as stronger reasoning or as prompt-time scaffolding. A topically related skill may still introduce an irrelevant convention, a misleading caveat, or a partially matched algebraic template. Such cases are easy to miss if retrieval is evaluated only by aggregate accuracy or lexical similarity. The distinction is the central motivation of this work: *retrieval relevance is not the same as skill applicability*. We therefore ask: when do TRS-style retrieved skills still provide useful scaffolding, and when do non-applicable skills become distractors?

We study this question through a controlled diagnostic evaluation of TRS-style skill-card retrieval on DeepMath-103K. We vary model capability by comparing Direct solving with a full TRS-style condition for GPT-4o-mini, GPT-5.4, and GPT-5.5, and we vary retrieval precision for the strongest solver by replacing the top-1 retrieved skill with rank-2, rank-5, or random skills. All main comparisons are paired by problem ID and use a two-stage evaluator with exact or symbolic checks followed by LLM adjudication for remaining cases. We also analyze harmful cases, BM25 score gaps, token-cost profiles, and lightweight mitigation attempts.

Figure 1 summarizes the diagnostic framing of our study.

[1]School of Mathematical Sciences; [2]Medical College; [3]Research Institute of Intelligent Complex Systems, Fudan University, Shanghai 200433, China. Correspondence to: Ying Wang <wangying_fudan@fudan.edu.cn>.

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).

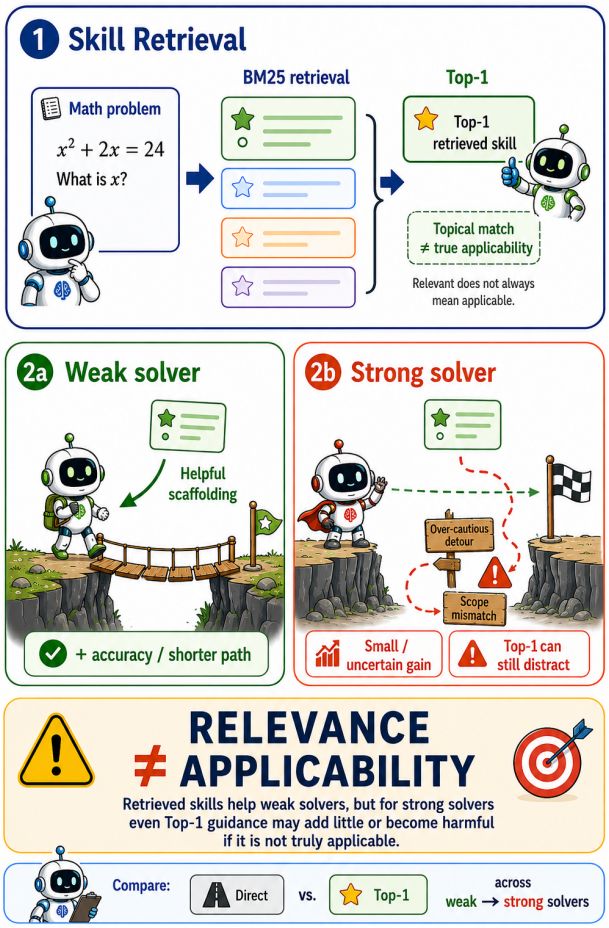

Figure 1. Overview of our diagnostic study. We compare Direct, top-1, off-rank, and random skill conditions across solver capability, using paired helpful, harmful, and distractor outcomes to separate useful scaffolding from inapplicable retrieved guidance.

The results show a clear model-dependent pattern. In the unmasked TRS-style setting, skill-card retrieval gives large scaffolding gains for weaker solvers, but only a small and uncertain gain for GPT-5.5, the strongest solver in our study. Exact source masking gives the sharper control: after same question source cards are removed, the unmasked gains do not persist, and GPT-5.5 falls below Direct. Lower-ranked and random replacements shift paired flips toward harm, showing that the benefit is not explained by simply adding skill-like text. Mechanism audits show that harmful cards are often partially applicable rather than irrelevant: they introduce over-cautious qualifications, convention shifts, wrong limiting regimes, or symbolic drift. BM25 score gaps help identify weak off-rank distractors, but do not reliably calibrate top-1 applicability; warning-style retrieval, unsupervised gating, and skeptical prompting likewise do not remove the distractor effect.

This paper contributes an evaluation protocol and diagnostic framework for skill-card retrieval, organized around the relevance–applicability distinction.

- We introduce a paired evaluation protocol for TRS-style skill-card retrieval across a model-capability gradient and retrieval-rank perturbations, separating scaffolding gains for weaker solvers from marginal guidance for strong solvers.
- We operationalize skill applicability through helpful, harmful, and off-rank distractor outcomes, making the distinction between retrieved relevance and mathematical applicability measurable.
- We audit distractor mechanisms and BM25 score features, showing that topical relevance can induce caveats, scope mismatch, convention shifts, or symbolic drift while remaining insufficient for top-1 applicability calibration.

2. Related Work

Mathematical reasoning models and evaluation. MATH and GSM8K standardized natural-language mathematical problem solving (Hendrycks et al., 2021; Cobbe et al., 2021), while chain-of-thought, self-consistency, program-of-thought prompting, Llemma, and MAmmoTH strengthened common test-time and model-training baselines (Wei et al., 2022; Wang et al., 2023; Chen et al., 2023; Azerbayev et al., 2024; Yue et al., 2024). Formal reasoning provides complementary signals through proof-assistant interaction, as in LeanDojo, DeepSeek-Prover-V1.5, and PUTNAMBENCH (Yang et al., 2023; Xin et al., 2025; Tsoukalas et al., 2024). Other benchmarks broaden the surface to multimodal reasoning, retrieval, and optimization modeling (Lu et al., 2024; Alshammari et al., 2026; Yang et al., 2025). Our setting remains natural-language problem solving with answer-level evaluation: we do not study proof-assistant premise retrieval, but retrieved procedural guidance for an informal solver.

Retrieval for mathematical reasoning. Retrieval-augmented generation and dense passage retrieval show how external text can serve as non-parametric context for language models (Lewis et al., 2020; Karpukhin et al., 2020). In mathematical domains, retrieval selects premises, lemmas, examples, or structurally related neighbors (Irving et al., 2016; Yang et al., 2023; Alshammari et al., 2026). The closest prior work to ours is TRS, where Zhao et al. (2026) retrieve compact skill cards distilled from prior reasoning trajectories and use them as procedural guidance rather than raw examples. We build on this view, but ask how skill benefit changes with model strength and how lower-ranked or random cards affect a strong solver.

Efficient test-time reasoning. Efficient test-time reasoning methods try to reduce unnecessary deliberation while preserving accuracy. TRS uses reusable guidance to shorten completions (Zhao et al., 2026); Accordion-Thinking compresses long reasoning histories with self-regulated summaries (Yang et al., 2026); and RLVR-style work studies performance and diversity under verifiable rewards and repeated sampling (Liang et al., 2026). Our analysis is complementary: skill cards add prompt tokens but may shorten completions, and off-rank cards can reduce accuracy while increasing total token use. Efficiency claims for retrieved reasoning should therefore be tied to retrieval precision and applicability.

Positioning. Our work lies at the intersection of retrieval-augmented mathematical problem solving and efficient test-time scaffolding. It is not a new theorem prover, benchmark, or full TRS reproduction; it is a controlled robustness study of TRS-style skill-card retrieval. We focus on the distinction between retrieval relevance and skill applicability, which becomes more important as the marginal benefit of additional guidance shrinks for stronger reasoners.

3. Problem Setup

We study test-time mathematical problem solving with a fixed external library of skill cards. In the direct condition, the solver receives only the problem prompt and produces a final answer. In the skill-augmented condition, the system first retrieves a card from the library, inserts the retrieved card into the prompt, and then calls the same base solver. A skill card is intended to be a compact procedural abstraction: it gives cues for when a reasoning pattern may apply, a short recipe for using it, and potential pitfalls. In this sense, skill-card retrieval differs from retrieving full worked examples or proof-assistant premises. The retrieved item is advice, and its value depends on whether that advice is applicable to the current problem.

In this paper, *TRS-style* refers to the online retrieval mechanism used in Thinking with Reasoning Skills (Zhao et al., 2026): a fixed skill-card bank, TRS-aligned retrieval text, lexical BM25 ranking (Robertson & Zaragoza, 2009), and prompt injection before solving. We use BM25 because it is the retrieval mechanism in the released TRS setting, not because it is meant to represent all possible modern retrievers. We use released TRS skill cards as the external library. We do not train the solver, generate new skill cards from our evaluation set, update card utilities, or reimplement the offline skill-construction stage of TRS. This lets us isolate the test-time behavior of retrieved skills while keeping the base model and target problems fixed across conditions.

Our robustness study varies which card is supplied to the solver. The *top-1* condition uses the highest-ranked BM25 card for the problem. The *rank-2* and *rank-5* conditions replace it with lower-ranked cards retrieved for the same problem, preserving some lexical relation to the query while weakening retrieval precision. The *random* condition supplies a card sampled independently of the target problem. Together with the direct baseline, these conditions separate the effect of adding skill-like text from the effect of retrieving an actually useful skill.

The central distinction is between retrieval relevance and skill applicability. Retrieval relevance is induced by the retriever: a card may share surface terms or a broad topic with the problem. Skill applicability is mathematical: following the card should be valid for the specific objects, conventions, and conditions in the problem. A retrieved card can therefore be relevant but non-applicable, for example by introducing an unnecessary caveat or shifting the scope of the answer. In our audit, a leading term card helps a complex analysis problem by selecting the correct large arc argument, while a Borel hierarchy card hurts by adding a caveat that answers a stricter question than the benchmark asks.

We use the following operational definitions throughout the results and case-card audit. The outcome labels are paired: they compare two conditions on the same problem rather than assigning an absolute quality label to a problem or card.

Paired Applicability Outcomes

Retrieval relevance. A lexical or topical match induced by the retriever, measured in our experiments by BM25 ranking and score features.

Skill applicability. The retrieved card can be followed without violating the problem's objects, assumptions, conventions, limiting regime, or requested answer form.

Helpful. Direct is incorrect on a problem, while the skill-augmented condition is correct on the same problem.

Harmful. Direct is correct, while the skill-augmented condition is incorrect on the same problem.

Off-rank distractor. Top-1 retrieval is correct, while a rank-2, rank-5, or random replacement skill is incorrect on the same problem.

These paired definitions are important because the strongest solver already has high direct accuracy. Small aggregate differences can correspond to meaningful shifts in which problems are solved or broken by retrieved guidance, so we analyze both accuracy and discordant paired outcomes.

4. Experimental Setup

Data, skill bank, and models. We evaluate on a locally processed subset of the DeepMath-103K test split (He et al., 2025). The local target pool contains 2000 problems selected with a fixed seed and split into 1000 build rows and 1000 evaluation rows. The main distractor-robustness evaluation uses the 1000-problem evaluation split, while the

model-capability gradient uses the first aligned 500 problems from that split so that each solver is compared under the same direct and skill-augmented conditions. The build rows are not used to generate new skill cards; they are used only for negative failure-memory experiments discussed later.

For retrieval, we use the released TRS DeepMath skill-card archive as a fixed skill-card bank (Zhao et al., 2026). This choice follows the intended TRS-style setting: skills are distilled from DeepMath problems and then reused as procedural guidance on related problems. In our runner, the solver prompt receives only the formatted skill text, not the stored answer field from the archive. At the same time, a source-overlap check found that the released archive contains a source-question match for every problem in our local DeepMath target pool. We therefore do not present the main experiments as a contamination-free generalization benchmark. Instead, we use them as a diagnostic evaluation of test-time skill use under the released TRS-style skill-card setting. The unmasked gains should be read in that setting, not as evidence of generalization to unseen problems. To quantify sensitivity to direct same-question cards, we also run an exact-source-masked control that removes every card whose normalized source question exactly matches the target before BM25 selection. The unmasked top-1 card is sourced from the exact target problem for 768/1000 problems in the main evaluation and 388/500 in the aligned subset; Appendix A reports the full overlap and retrieval-only statistics.

The main strong-solver experiments use the API model alias GPT-5.5; to measure how retrieval benefit changes with model capability, we additionally evaluate GPT-4o-mini and GPT-5.4 on the same 500-problem subset. All models are queried through the same OpenAI-compatible chat-completions runner, and the base model is not fine-tuned or updated.

Retrieval conditions. We run two families of conditions. The model-capability experiment compares *Direct* solving, *Full TRS-style* retrieval, and *Exact-source-masked top-1* retrieval for each solver on the same aligned 500 problems. Direct receives only the problem prompt, while both retrieval conditions use the same TRS-style hint wrapper; the masked condition differs only by removing exact source-question cards before retrieval. The distractor-robustness experiment focuses on GPT-5.5 and compares five matched conditions on 1000 problems: Direct, top-1 skill retrieval, rank-2 skill retrieval, rank-5 skill retrieval, and random-skill retrieval. The off-rank and random conditions are not intended as useful systems. They are controlled perturbations that test whether performance depends on retrieving an applicable skill rather than simply adding skill-like text.

Sampling checks. To check that the main subsets are not cherry-picked with respect to retrieval difficulty, we also run a retrieval-only audit over the local 2000-problem pool, the 1000-problem evaluation split, and the aligned 500-problem subset. This audit makes no solver calls. The 1000 and 500 subsets have similar median top-1 BM25 scores, score gaps, problem lengths, retrieved-card lengths, and top-1 topic-prefix distributions to the local pool; the topic total-variation distances from the pool are 0.028 and 0.045, respectively. Full statistics are reported in Appendix A.

Evaluation and statistics. We report adjudicated accuracy as the primary metric. A rule-based evaluator first accepts high-confidence exact or equivalent matches, including TRS-style exact matching, math-verify checks, symbolic equivalence through SymPy when applicable, and conservative handling of boolean or multiple-choice labels. Remaining rule-false or risky cases are sent to an LLM adjudicator that compares the model's final answer against the reference answer. We keep strict rule-based accuracy for auditability, but use adjudicated accuracy for the main analysis because exact string matching can underestimate mathematically equivalent answers. Technical failures such as API errors, empty responses, or token exhaustion are not counted as mathematical errors. Uncertain adjudicator verdicts are not assigned a default label; they are resolved through targeted second-pass adjudication and, where necessary, manual inspection. Final reported results use rerun and readjudicated clean files with no API errors, unevaluated rows, or uncertain verdicts.

Because the main strong solver has high direct accuracy, we use paired statistics throughout. For each skill condition, we compute helpful and harmful counts relative to Direct on the same problem set, report paired accuracy differences with bootstrap confidence intervals (Efron, 1979), and use McNemar's exact test for paired significance (McNemar, 1947). Bootstrap confidence intervals use 10,000 resamples of problem IDs with replacement and recompute the paired accuracy difference; they quantify problem-level uncertainty over fixed outputs and fixed adjudication, not adjudicator noise. Per-condition Wilson confidence intervals are reported in Appendix B (Wilson, 1927). We also record prompt tokens, completion tokens, total tokens, and latency. Token measurements separate prompt inflation from completion-token reduction, and normalized proxy costs are reported only under a fixed output-token-heavy weighting for within-experiment comparison.

Implementation details. All solver calls use temperature 0, a single user-message chat prompt, and the same answer-extraction and adjudication code across conditions. The distractor-robustness runner uses up to 8192 completion tokens, deterministic random-skill selection with a fixed seed,

Table 1. Model-capability gradient and exact-source-masked control on an aligned 500-problem subset. Accuracies are percentages, and deltas are paired against Direct with 95% problem-level bootstrap CIs. H/H gives helpful/harmful counts relative to Direct, and p columns report two-sided exact McNemar tests against Direct. Wilson CIs for standalone accuracies and full statistics are in Appendix B.

Model	Direct	TRS-style	Masked	Δ_{TRS}	Δ_{mask}	H/H_{TRS}	H/H_{mask}	p_{TRS}	p_{mask}
GPT-4o-mini	60.4	70.8	59.2	+10.4 [+6.4,+14.6]	-1.2 [-5.0,+2.6]	84/32	45/51	$< .001$	.610
GPT-5.4	87.6	91.8	86.2	+4.2 [+1.4,+7.0]	-1.4 [-3.8,+1.0]	37/16	17/24	.005	.349
GPT-5.5	93.0	93.6	90.4	+0.6 [-0.8,+2.2]	-2.6 [-4.8,-0.4]	9/6	9/22	.607	.029

and up to three API retry attempts. Reruns only replace technical failures or uncertain adjudications, not mathematical errors. We do not override provider-side model snapshots or routing, so the reported model names should be read as API aliases rather than immutable open checkpoints. For auditability, we retain records needed to reproduce the paired comparisons and case-card labels.

5. Results

Skill-card scaffolding benefit decreases with model capability. Table 1 compares Direct solving, Full TRS-style retrieval, and exact-source-masked top-1 retrieval on the same 500-problem subset for three solvers. In the unmasked TRS-style setting, the benefit is strongly model-dependent. GPT-4o-mini gains 10.4 percentage points from skill-card retrieval, and GPT-5.4 gains 4.2 points. For GPT-5.5, however, the lift is only 0.6 points, with a paired confidence interval crossing zero. The paired helpful and harmful counts show the same pattern: weaker models have many more problems where the retrieved skill repairs a direct failure, while the strongest model has little remaining headroom for positive guidance.

The exact-source-masked control gives the complementary interpretation. After removing same-question source cards before retrieval, the unmasked gains do not persist: GPT-4o-mini changes by -1.2 points relative to Direct, GPT-5.4 by -1.4 points, and GPT-5.5 by -2.6 points. The first two masked deltas are not significant, while the GPT-5.5 masked condition is significantly below Direct. This does not invalidate the TRS-style setting, since TRS explicitly studies reuse of distilled skills from the archive. It does show that the apparent teacher effect in this release is sensitive to exact source-card availability, and it sharpens the diagnostic question we study: once direct same-question cards are removed, lexical relevance alone does not guarantee an applicable skill.

For strong solvers, retrieval precision matters more than adding a skill. We next isolate retrieval precision on GPT-5.5 using the 1000-problem distractor-robustness evaluation. Table 2 shows that top-1 retrieval changes accuracy only slightly, from 93.4% to 94.0% (+0.6 pp; paired 95% bootstrap CI -0.4 to $+1.7$), with 17 helpful and 11 harmful paired flips. We therefore do not treat the top-1 gain over Direct as a reliable aggregate improvement. By contrast, rank-5 retrieval is worse than Direct (-1.1 pp; paired 95% CI -2.0 to -0.2; McNemar exact $p = .035$), and is also worse than top-1 retrieval. These results argue against a generic "skill text helps" explanation: the top-ranked skill is useful only when retrieval selects sufficiently applicable guidance.

Table 2. Distractor-robustness results on 1000 GPT-5.5 problems. Deltas are paired against Direct. Confidence intervals are problem-level paired bootstrap CIs. H/H and Net are paired against Direct, and p-values are two-sided exact McNemar tests over discordant pairs. Wilson CIs for standalone accuracies and additional pairwise comparisons are in Appendix B.

Condition	Acc.	Δ [95% CI]	H/H	Net	p	Tokens
Direct	93.4	–	–	–	–	1138
Top-1	94.0	+0.6 [-0.4,+1.7]	17/11	+6	.345	1379
Rank-2	93.0	-0.4 [-1.4,+0.6]	10/14	-4	.541	1457
Rank-5	92.3	-1.1 [-2.0,-0.2]	6/17	-11	.035	1478
Random	92.4	-1.0 [-2.0,+0.0]	8/18	-10	.076	1434

The paired counts clarify why aggregate accuracy alone is insufficient. Top-1 retrieval repairs 17 direct failures but also breaks 11 direct successes. Rank-5 and random retrieval have the opposite balance, with fewer helpful cases and more harmful cases. Because the Direct baseline is already strong, these discordant cases determine the net effect. The key result is therefore not that retrieval always helps or hurts, but that retrieval precision controls whether skill-card prompting is beneficial.

A neutral-length control does not reproduce the top-1 gain. To separate skill content from prompt length, we also run a 500-problem control that inserts non-mathematical neutral text matched to the top-1 skill length inside the same hint wrapper. On the same 500 problem IDs, Direct reaches 92.0%, neutral-length context reaches 91.2%, and top-1 retrieval reaches 93.2%. The neutral condition is not significantly different from Direct (-0.8 points; 95% CI $[-2.4, 0.8]$; $p = 0.4807$), while top-1 retrieval is significantly better than neutral context ($+2.0$ points; 95% CI $[0.6, 3.6]$; $p = 0.0213$). This weakens a pure context-length explanation for the top-1 benefit. At the same time, neutral context shortens completions and changes token usage, so length and prompt style remain measurable controls rather

Table 3. Manual-audit taxonomy of harmful cases. Counts are primary labels and rows sum to Harmful after adding Other. Cautious denotes over-cautious qualification; Scope covers scope, precondition, and convention mismatch; Drift covers numeric or symbolic drift.

Condition	Harm.	Cautious	Scope	Drift	Other	ΔTok.
Top-1	11	8	1	1	1	+224
Rank-2	14	5	0	5	4	+1056
Rank-5	17	6	1	6	4	+586
Random	18	7	0	5	6	+505

than irrelevant nuisances.

Skill retrieval changes the token profile. Skill-card prompting adds context, so all skill conditions use more total tokens than Direct in the GPT-5.5 distractor experiment. Top-1 retrieval increases average total tokens by 21.2% relative to Direct, but shortens completions by 19.8%. Under a fixed output-heavy proxy that weights completion tokens more than prompt tokens, this corresponds to a 12.2% lower normalized proxy cost; under equal-token accounting, it is more expensive. The result is therefore not a deployment-level cost claim, but a token-profile shift: applicable top-1 skills can shorten completions, while off-rank and random skills are worse than top-1 on both accuracy and total-token use. Appendix D reports the prompt, completion, total-token, and proxy-index breakdowns.

Overall, the main results show a shift in bottleneck as solver capability increases. For weaker solvers, retrieved skills provide substantial positive guidance. For the strongest solver, the average top-1 gain is small, and the main question becomes whether the retrieved skill is applicable. The next section analyzes the failure modes behind this distractor effect.

6. Analysis: When Retrieved Skills Become Distractors

The main results show that retrieval precision matters, but they do not explain how a retrieved skill changes a solver's answer. We therefore audit harmful cases in the GPT-5.5 distractor-robustness run. This audit is diagnostic: it does not change the evaluator or the reported accuracy numbers. For each condition, we examine cases where Direct is correct and the skill condition is wrong, and cases where top-1 retrieval is correct but an off-rank or random replacement is wrong. We assign one primary mechanism label and optional secondary labels; Table 3 reports primary labels grouped into common categories, with residual cases placed in Other.

Harm often comes from partial applicability, not complete irrelevance. Table 3 summarizes the harmful cases relative to Direct. The largest recurring pattern is over-cautious or conditional answering: the retrieved skill leads the solver to add a caveat, distinguish cases not requested by the problem, or answer only under an extra condition. This is especially visible for top-1 harmful cases, where 8 of 11 cases fall into this category. Lower-ranked and random skills more often induce numeric or symbolic drift, consistent with the view that they push the solver onto a longer and less stable reasoning path. These cases are rare in absolute terms, but they matter because the strong direct solver leaves little headroom: on the main GPT-5.5 run, top-1 retrieval has 17 helpful cases and 11 harmful cases.

Skill Applicability Case Cards. The following case cards use a fixed audit format. They are not additional evaluation examples; they are mechanism samples drawn from the paired discordances in the clean GPT-5.5 run. The trace fields quote or paraphrase visible answer text, not hidden reasoning. Case IDs encode the paired outcome type: Helpful-k denotes Direct wrong and top-1 retrieval correct, Harmful-k denotes Direct correct and top-1 retrieval wrong, and Distractor-k denotes a lower-rank or random replacement that breaks a top-1 success.

Helpful-1: Helpful top-1 skill

Case ID / domain. Helpful-1, complex analysis.

Problem requirement. For $p(z) = z^4 + z^3 + 4z^2 + 3z + 7$, determine the change in argument along a sufficiently large quarter-circle arc C_3.

Retrieved skill. Use leading-term behavior on large arcs; for a degree-d polynomial, $\Delta \arg p(z) \approx d\,\Delta\theta$.

Key answer trace. Direct follows the supplied $\arg(p(iy))$ expression and outputs 0. Top-1 retrieval instead uses $p(z) \sim z^4$ and returns $4 \cdot \pi/2 = 2\pi$.

Mathematical diagnosis. The direct solution tracks an axis parametrization rather than the requested large circular arc. The retrieved card is applicable because on C_3, z^4 dominates the lower-order terms, so the argument winds four times as fast as z.

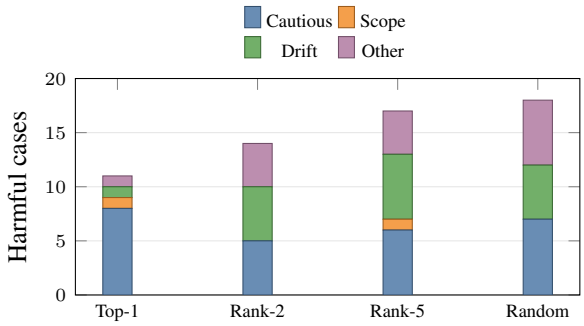

Figure 2. Primary mechanism labels for harmful cases. Top-1 harms are dominated by over-cautious qualification, while lower-ranked and random cards induce more numeric or symbolic drift.

Mechanism label. Helpful: representation choice / theorem selection.

Harmful-1: Harmful top-1 skill

Case ID / domain. Harmful-1, descriptive set theory.

Problem requirement. Answer whether every Borel set can be constructed from open sets using countable unions, intersections, and complements.

Retrieved skill. Use the Borel hierarchy and be careful about transfinite countable-ordinal stages.

Key answer trace. Direct answers yes. The skill condition answers "not in general by a single countable sequence" and adds that some Borel sets require countable-ordinal stages.

Mathematical diagnosis. The caveat is mathematically relevant to construction depth, but it answers a stricter question than the benchmark asks. Every Borel set belongs to the σ-algebra generated by open sets; the transfinite hierarchy refines how many stages may be needed, not whether closure under countable operations generates the set.

Mechanism label. Harmful: over-cautious qualification / scope mismatch.

Harmful-2: Harmful top-1 skill

Case ID / domain. Harmful-2, enumerative combinatorics.

Problem requirement. Count permutation-style colorings of a 5×5 grid when two corner cells are excluded; the benchmark answer is 78.

Retrieved skill. Map row-column selections to permutations and apply inclusion–exclusion for forbidden cells.

Key answer trace. The skill condition returns 72 for adjacent corners and 78 for opposite corners, rather than giving the benchmark's single answer 78.

Mathematical diagnosis. The permutation model is relevant: one excluded corner removes $4!$ permutations, and opposite excluded corners add back $3!$. The failure is not arithmetic but answer-form drift: the retrieved card encourages a case distinction about corner geometry that changes a single-answer benchmark item into a conditional response.

Mechanism label. Harmful: over-cautious qualification / convention shift.

Distractor-1: Off-rank distractor

Case ID / domain. Distractor-1, complex analysis.

Problem requirement. Using Rouché's theorem, determine the number of zeros of $1 + z + z^2/2! + \cdots + z^n/n!$ inside $|z| = R$.

Retrieved skill. A lower-ranked Rouché-style card suggests comparing the partial sum with e^z on a fixed circle.

Key answer trace. The replacement condition writes $e^z - P_n(z) = \sum_{k=n+1}^{\infty} z^k/k!$, takes $n \to \infty$ with fixed R, and concludes 0 zeros.

Mathematical diagnosis. That comparison can be meaningful for a different asymptotic question, but it changes the benchmark's limiting regime. The intended Rouché argument compares the leading polynomial term $z^n/n!$ with lower-order terms on a sufficiently large contour, giving n zeros.

Mechanism label. Off-rank distractor: wrong limiting regime / template overfitting.

Lexical score gaps identify weak distractors but not top-1 applicability. We next ask whether BM25 scores can predict harmful retrieval. The answer is mixed. Top-1 harmful cases do not have systematically lower top-1 scores or smaller top1–rank2 gaps than benign top-1 cases: the median top1–rank2 gap is 9.83 for top-1 harmful cases and 10.71 for cases where both Direct and top-1 are correct, and the comparison probability is close to random ($P(\text{event} < \text{control}) = 0.49$). In contrast, score gaps are informative for off-rank distractors. When rank-2 is wrong but top-1 is correct, the median top1–rank2 gap is 31.93 rather than 10.71; when rank-5 is wrong but top-1 is correct, the median top1–rank5 gap is 38.00 rather than 20.30. Thus, BM25 is useful as a ranking signal that separates better and worse retrieved cards, but it is not a calibrated applicability signal for deciding whether the top-1 card should be used.

Overall, the analysis supports the distinction between relevance and applicability. Lexical retrieval can often identify cards that are broadly related to the problem, but strong solvers are sensitive to whether the retrieved procedure matches the exact mathematical conditions of the task. The resulting errors are not random noise; they are structured distractor effects caused by partially applicable guidance.

7. Scope and Negative Attempts

Our study is a diagnostic evaluation rather than a full TRS reproduction or a contamination-free benchmark. We use the released DeepMath skill-card archive as a fixed retrieval bank, which isolates online retrieval and test-time applicability but inherits source-question overlap with the target distribution. To keep this scope explicit, we expose only skill text rather than stored answers, add exact-source-masked controls, and report paired comparisons on fixed problem sets. The conclusions should therefore be read as evidence about applicability under this TRS-style setting, not as universal claims over all mathematical tasks, future API snapshots, or formally verified proofs. The strongest solver result is also limited to one API model alias, GPT-5.5, and the qualitative taxonomy is based on the observed harmful flips rather than a large independent annotation study.

We also tested several lightweight remedies motivated by the analysis. Table 4 summarizes them. Failure memory retrieves warning-style cards distilled from previous failures rather than positive skills, but the warnings were often generic and did not outperform the main TRS-style condition. A retrospective LLM gate was too conservative, selecting few skills and missing helpful cases; by contrast, an oracle gate indicates that selective use could help if ap-

Table 4. Negative mitigation attempts. FM uses the 1000-problem held-out evaluation. Gate uses a 500-problem held-out subset. Prompt calibration is evaluated on 23 selected discordant cases, so its accuracy is not comparable to full-distribution accuracy.

Attempt	Setting	Selection / retention	Acc. (%)	Main failure mode
Failure memory	1000 eval	6 helpful / 10 harmful vs Direct	94.43	Generic warnings; concentrated retrieval
LLM gate	500 eval	Uses TRS on 18/500; captures 0/7 helpful	95.17	Too conservative; high gate-token cost
Oracle gate	500 eval	Uses TRS on 7/500	96.54	Upper bound, not deployable
Cautious prompts	23 discordant	Retains 10/16 helpful; recovers 1/7 harmful	47.83–50.00	Drops helpful cases

plicability were known. Skeptical and applicability-check prompts reduced some harms but also discarded useful skills. These negative attempts are useful mainly as controls: they suggest that the bottleneck is not solved by adding caution, but requires a sharper applicability signal.

8. Conclusion

We studied whether TRS-style skill-card retrieval strengthens mathematical reasoning or mainly provides test-time scaffolding. Across a model-capability gradient, retrieved skills provide substantial gains for weaker solvers but only a small marginal gain for the strongest solver we evaluate. In the strong-solver setting, retrieval precision becomes critical: the top-1 retrieved skill slightly improves over direct solving, while lower-ranked and random skills can reduce accuracy and increase total token use. Our analysis shows that harmful retrieval is often not complete irrelevance, but partial applicability: a skill can be topically related while introducing an unnecessary caveat, convention shift, or unstable symbolic path.

These findings suggest a shift in how skill-card systems should be evaluated for strong mathematical reasoners. The question is not only whether retrieved skills contain useful mathematical experience, but whether the system can decide when that experience applies. Lexical retrieval scores help rank candidate cards, but they are not sufficient applicability signals for the top-ranked card, and simple warning or gating strategies do not reliably remove the distractor effect. Future systems must show applicability-aware selective use before skill-card retrieval can be interpreted as reasoning improvement rather than prompt-time scaffolding.

Impact Statement

This work studies evaluation and reliability for mathematical reasoning systems that use retrieved procedural guidance. The main societal risks are the same as for other automated mathematical assistants: users may over-trust fluent but invalid reasoning, especially when retrieved context makes the answer appear more grounded. Our diagnostic focus is intended to reduce that risk by making applicability failures easier to measure and audit.

References

Alshammari, S., Wen, K., Zainal, A., Hamilton, M., Safaei, N., Albarakati, S., Freeman, W. T., and Torralba, A. MathNet: A global multimodal benchmark for mathematical reasoning and retrieval. In *International Conference on Learning Representations*, 2026.

Azerbayev, Z., Schoelkopf, H., Paster, K., Dos Santos, M., McAleer, S. M., Jiang, A. Q., Deng, J., Biderman, S., and Welleck, S. Llemma: An open language model for mathematics. In *International Conference on Learning Representations*, 2024.

Chen, W., Ma, X., Wang, X., and Cohen, W. W. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Transactions on Machine Learning Research*, 2023.

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

Efron, B. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

He, Z., Liang, T., Xu, J., Liu, Q., Chen, X., Wang, Y., Song, L., Yu, D., Liang, Z., Wang, W., Zhang, Z., Wang, R., Tu, Z., Mi, H., and Yu, D. DeepMath-103K: A large-scale, challenging, decontaminated, and verifiable mathematical dataset for advancing reasoning, 2025. arXiv preprint arXiv:2504.11456.

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. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Irving, G., Szegedy, C., Alemi, A. A., and Eén, N. DeepMath: Deep sequence models for premise selection. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W.-t. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., and Kiela, D. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Liang, X., Li, Z., Gong, Y., Shen, Y., Wu, Y. N., Guo, Z., and Chen, W. Beyond Pass@1: Self-play with variational problem synthesis sustains RLVR. In *International Conference on Learning Representations*, 2026.

Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.-W., Galley, M., and Gao, J. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations*, 2024.

McNemar, Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

Robertson, S. and Zaragoza, H. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009.

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. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0368.

Wang, X., Wei, J., Schuurmans, D., Le, Q. V., Chi, E. H., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E. H., Le, Q. V., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Wilson, E. B. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

Xin, H., Ren, Z. Z., Song, J., Shao, Z., Zhao, W., Wang, H., Liu, B., Zhang, L., Lu, X., Du, Q., Gao, W., Zhu, Q., Yang, D., Gou, Z., Wu, Z. F., Luo, F., and Ruan, C. DeepSeek-Prover-V1.5: Harnessing proof assistant feedback for reinforcement learning and monte-carlo tree search. In *International Conference on Learning Representations*, 2025.

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. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Yang, Z., Wang, Y., Huang, Y., Guo, Z., Shi, W., Han, X., Feng, L., Song, L., Liang, X., and Tang, J. OptiBench meets ReSocratic: Measure and improve LLMs for optimization modeling. In *International Conference on Learning Representations*, 2025.

Yang, Z., Guo, Z., Huang, Y., Wang, Y., Shi, W., Wang, Y., Liang, X., and Tang, J. Accordion-thinking: Self-regulated step summaries for efficient and readable LLM reasoning, 2026. arXiv preprint arXiv:2602.03249.

Yue, X., Qu, X., Zhang, G., Fu, Y., Huang, W., Sun, H., Su, Y., and Chen, W. MAmmoTH: Building math generalist models through hybrid instruction tuning. In *International Conference on Learning Representations*, 2024.

Zhao, G., Shi, Q., Xiao, X., Zhang, X., Yang, T., and Sun, L. Thinking with reasoning skills: Fewer tokens, more accuracy, 2026. arXiv preprint arXiv:2604.21764.

A. Subset Representativeness and Source Overlap

Table 5 reports retrieval-only statistics for the local 2000-problem pool, the 1000-problem evaluation split, and the aligned 500-problem subset used in the model-capability gradient. No solver calls are made in this audit. Topic total-variation distance uses the top-1 retrieved-card topic prefix as a proxy because the target rows do not carry complete domain labels.

Table 5. Retrieval-only subset representativeness. Scores are BM25 values. Gap12 is top1–rank2; gap15 is top1–rank5.

Set	n	Top1	Gap12	Gap15	Problem tok.	Skill tok.	Topic TVD	Top1 exact source
Local-2000 pool	2000	73.73	11.21	21.15	33	253	0.000	1600/2000
Main-1000 eval	1000	73.18	11.31	21.26	33	254	0.028	768/1000
Aligned-500	500	73.45	12.35	22.45	33	255	0.045	388/500

We also compare evaluation problems against the source question and retrieval text stored in the released skill-card archive. The archive contains an exact source-question match for every problem in both local target files: 2000/2000 in the local pool and 1000/1000 in the evaluation split. Although the raw archive stores source answers, our retrieval document uses the source question, topic, and skill text, and the solver prompt injects only the formatted skill text. Table 6 reports the retrieval-only effect of masking exact source-question cards before selecting BM25 results. On the aligned 500-problem subset, 388/500 unmasked top-1 cards are exact source matches; after masking, the median selected card has original BM25 rank 2 and median BM25 score 54.99 rather than 73.45. This confirms that the masked solver control removes direct same-question card use, while still leaving highly related same-distribution cards available.

Table 6. Retrieval-only exact-source masking audit. Original rank is the unmasked BM25 rank of the selected masked top-1 card.

Set	n	Exact source	Orig. top1 exact	Top1 changed	Orig. top1	Masked top1	Orig. rank
Local-2000 pool	2000	2000/2000	1600/2000	1600/2000	73.73	57.04	2
Main-1000 eval	1000	1000/1000	768/1000	768/1000	73.18	56.18	2
Aligned-500	500	500/500	388/500	388/500	73.45	54.99	2

B. Full Statistical Tables

The main tables report paired deltas because all comparisons are matched by problem ID. This appendix adds the standalone Wilson intervals and pairwise tests that were omitted from the main text to avoid duplicating the primary result tables. The 1000-problem GPT-5.5 run has no remaining API errors, unevaluated rows, or uncertain adjudications after technical reruns.

Table 7. Supplementary statistics for Table 1. The first three columns give standalone Wilson 95% confidence intervals for adjudicated accuracy. The final columns report the pairwise masked-vs-unmasked comparison omitted from the main table.

Model	Direct Wilson CI	TRS Wilson CI	Masked Wilson CI	Masked − TRS	p
GPT-4o-mini	[56.0,64.6]	[66.7,74.6]	[54.8,63.4]	-11.6 [-15.6,-7.8]	$< .001$
GPT-5.4	[84.4,90.2]	[89.1,93.9]	[82.9,88.9]	-5.6 [-8.2,-3.2]	$< .001$
GPT-5.5	[90.4,94.9]	[91.1,95.4]	[87.5,92.7]	-3.2 [-5.6,-1.0]	.011

C. Neutral-Length Context Control

We add a 500-problem control that injects neutral, non-mathematical text with length matched to the top-1 skill card inside the same hint wrapper. The control reuses the same 500 problem IDs as the existing Direct, top-1, rank-5, and random outputs. Table 9 reports the matched comparison.

D. Token Breakdowns and Proxy Cost Index

Table 10 separates prompt and completion tokens. The proxy index uses a fixed output-heavy weighting in which completion tokens count six times as much as prompt tokens, and normalizes Direct to 100. It is included only for within-experiment comparison.

Table 8. Additional paired comparisons for the 1000-problem GPT-5.5 distractor run. Candidate-only means the candidate condition is correct and baseline is wrong; baseline-only means the reverse.

Comparison	n	Δ	95% CI	Cand.-only	Base-only	p
Top-1 − Direct	1000	+0.006	[-0.004,+0.017]	17	11	.3449
Rank-2 − Direct	1000	-0.004	[-0.014,+0.006]	10	14	.5413
Rank-5 − Direct	1000	-0.011	[-0.020,-0.002]	6	17	.0347
Random − Direct	1000	-0.010	[-0.020,+0.000]	8	18	.0755
Rank-2 − Top-1	1000	-0.010	[-0.021,+0.001]	10	20	.0987
Rank-5 − Top-1	1000	-0.017	[-0.028,-0.006]	8	25	.0046
Random − Top-1	1000	-0.016	[-0.027,-0.006]	6	22	.0037

Table 9. Neutral-length control on the first 500 GPT-5.5 evaluation problems. Accuracies are adjudicated with Wilson 95% CIs. Paired deltas use Direct as the baseline except for the Top-1 − Neutral row.

Condition	Acc.	Total tok.	Helpful	Harmful	Paired result
Direct	92.0 [89.3,94.1]	1226	–	–	–
Neutral length	91.2 [88.4,93.4]	1625	7	11	-0.8 pp; $p = 0.4807$
Top-1	93.2 [90.7,95.1]	1444	10	4	$+1.2$ pp; $p = 0.1796$
Rank-5	90.2 [87.3,92.5]	1589	1	10	-1.8 pp; $p = 0.0117$
Random	91.6 [88.8,93.7]	1545	5	7	-0.4 pp; $p = 0.7744$
Top-1 − Neutral	–	–	13	3	$+2.0$ pp; $p = 0.0213$

E. Taxonomy Codebook

For each harmful or distractor case, we assign one primary mechanism label and optional secondary labels. Helpful labels include theorem selection, representation choice, asymptotic simplification, case decomposition, and algebraic stabilization. Harmful labels include over-cautious qualification, scope or precondition mismatch, convention shift, wrong limiting regime, template overfitting, numeric or symbolic drift, and other. Over-cautious qualification means that the retrieved card induces a conditional or caveated answer when the benchmark asks for a definite answer. Scope or precondition mismatch means that the card's theorem, object class, or assumptions do not match the problem. Convention shift means that the card moves the solver to a different but plausible interpretation of a benchmark convention. Wrong limiting regime means that the card changes the order of limits or asymptotic variable. Template overfitting means that the solver follows the retrieved procedural template despite a missing structural premise. Numeric or symbolic drift means that the skill-augmented reasoning path introduces an arithmetic, algebraic, or symbolic error not present in the direct solution.

F. Negative Mitigation Details

The negative mitigation runs are intended as diagnostic controls rather than final methods. Failure-memory retrieval uses the held-out 1000-problem evaluation and retrieves warning-style memory before solving. Retrospective gating uses a 500-problem subset and asks an LLM gate whether to use the skill after seeing the problem and retrieved card. Prompt calibration is evaluated only on selected discordant cases, so its accuracy is not distributional.

G. Case-Card Audit

This appendix gives a curated case-card audit for representative paired discordances. Case IDs encode the paired outcome type: Helpful-k denotes Direct wrong and top-1 retrieval correct, Harmful-k denotes Direct correct and top-1 retrieval wrong, and Distractor-k denotes a lower-rank or random replacement that breaks a top-1 success. The trace fields quote or paraphrase visible final-answer text, not hidden reasoning.

We do not display every discordant row as a card. During audit, we exclude examples whose main lesson would be a disputed gold answer, an under-specified problem statement, or an apparent adjudication artifact rather than a retrieval mechanism. The raw artifact files retain the full discordant lists and adjudicator rationales; the cards below are intended to make the mechanism labels readable and auditable.

Table 10. Token and normalized proxy-cost summary for the 1000-problem GPT-5.5 run.

Condition	Prompt	Completion	Total	Total Δ	Proxy index
Direct	124	1014	1138	–	100.0
Top-1	565	813	1379	+241	87.8
Rank-2	565	892	1457	+319	95.3
Rank-5	565	912	1478	+340	97.3
Random	538	896	1434	+296	95.3

Table 11. Mitigation diagnostics. Oracle gate uses the observed paired outcomes and is therefore an upper bound, not a deployable method.

Attempt	n	Skill used / retained	Helpful retained	Harm avoided	Result
Failure memory	1000	–	6	–	94.43% accuracy; below full TRS-style
LLM gate	500	18	0/7	2/2	95.17% accuracy; close to Direct
Oracle gate	500	7	7/7	2/2	96.54% upper bound
Cautious prompts	23	–	10/16	1/7	Retains helpful cases poorly

G.1. Curated Top-1 Harmful Cases

Harmful-1: top-1 harmful, problem 1816

Case ID / domain. Harmful-1; descriptive set theory.

Problem requirement. Answer whether every Borel set can be constructed from open sets using countable unions, intersections, and complements.

Retrieved skill. Use the Borel hierarchy and be careful about transfinite countable-ordinal stages.

Key answer trace. Direct answers yes. Top-1 retrieval answers "not in general by a single countable sequence" and then adds that every Borel set is obtained by iterating the operations through countable ordinals.

Mathematical diagnosis. The retrieved card is topically relevant, but it changes the target statement. The benchmark asks whether the generated σ-algebra contains all Borel sets; the hierarchy caveat concerns construction depth. The final answer therefore becomes over-qualified relative to the requested yes/no form.

Mechanism label. over-cautious qualification / scope mismatch.

Harmful-2: top-1 harmful, problem 33

Case ID / domain. Harmful-2; enumerative combinatorics.

Problem requirement. Count row-column permutation colorings of a 5×5 grid when two corner cells are excluded; the benchmark expects the single value 78.

Retrieved skill. Map valid colorings to permutations and use inclusion–exclusion for forbidden cells.

Key answer trace. Direct returns 78. Top-1 retrieval returns 72 for adjacent excluded corners and 78 for opposite excluded corners.

Mathematical diagnosis. The permutation model is applicable, but the retrieved procedure encourages a case split that is not part of the benchmark answer form. The error is not the inclusion–exclusion template itself; it is the shift from a single convention-dependent answer to a conditional response.

Mechanism label. over-cautious qualification / convention shift.

Harmful-3: top-1 harmful, problem 1312

Case ID / domain. Harmful-3; Riemannian geometry.

Problem requirement. For a complete Riemannian manifold, decide whether geodesic balls around a fixed point exhaust the manifold with nested closures.

Retrieved skill. Before using global metric coverage, check whether all points are path-connected to the base point.

Key answer trace. Direct gives the expected exhaustion statement under the usual connected-manifold reading. Top-1 retrieval answers "not necessarily" and only affirms the statement after adding a connectedness assumption.

Mathematical diagnosis. The connectivity check is mathematically sensible, but it shifts the convention being used. In the benchmark setting, Riemannian manifold is read with the connectedness convention needed for geodesic distance from x_0 to all points; the retrieved card pushes the solver to reject the statement under a broader convention.

Mechanism label. scope/precondition mismatch / convention shift.

Harmful-4: top-1 harmful, problem 1421

Case ID / domain. Harmful-4; triangle geometry.

Problem requirement. Given an altitude and two medians in triangle ABC, find the single length BC; the benchmark answer is 20.

Retrieved skill. Use coordinate geometry: place BC on the x-axis and translate altitude and median constraints into equations.

Key answer trace. Direct returns 20. Top-1 retrieval solves the coordinate equations but reports 4 or 20.

Mathematical diagnosis. The coordinate method is applicable, but the skill-guided solution does not finish the geometric filtering step. The extraneous algebraic branch is incompatible with the stated triangle configuration, so the final answer violates the requested single-value form.

Mechanism label. numeric/symbolic drift / answer-form drift.

G.2. Helpful Top-1 Samples

Helpful-1: helpful, problem 303

Case ID / domain. Helpful-1; complex analysis.

Problem requirement. For $p(z) = z^4 + z^3 + 4z^2 + 3z + 7$, determine the change in argument along a sufficiently large quarter-circle arc.

Retrieved skill. On a large circular arc, replace a polynomial by its leading term when computing the asymptotic argument change.

Key answer trace. Direct follows the supplied $p(iy)$ expression and returns 0. Top-1 retrieval uses $p(z) \sim z^4$ on the large arc and returns 2π.

Mathematical diagnosis. The card is applicable because on the large quarter circle the leading term dominates, so $\Delta \arg p(z)$ is approximately $4\Delta\theta = 2\pi$. It repairs a representation error in the direct solution.

Mechanism label. representation choice / asymptotic simplification.

Helpful-2: helpful, problem 626

Case ID / domain. Helpful-2; analytic geometry.

Problem requirement. Count integer-coordinate points inside the bounded region determined by an ellipse and a hyperbola.

Retrieved skill. For regions bounded by multiple conics, first locate intersections and identify the actual enclosed region before counting lattice points.

Key answer trace. Direct returns 0. Top-1 retrieval identifies the intended enclosed region and returns 13.

Mathematical diagnosis. The retrieved card supplies the missing geometric setup. It changes the solver from treating the conics as failing to enclose a relevant region into enumerating lattice points in the intended overlap.

Mechanism label. representation choice / case decomposition.

Helpful-3: helpful, problem 852

Case ID / domain. Helpful-3; discrete mathematics.

Problem requirement. Decide whether a chromatically rigid hypergraph has a minimal rigid edge subset.

Retrieved skill. For finite structures, use finite descent: any nonempty family of finite subsets has an inclusion-minimal member.

Key answer trace. Direct adds an infinite-edge caveat and does not give the benchmark's unconditional yes. Top-1 retrieval applies finite descent and answers yes.

Mathematical diagnosis. Under the finite-structure convention used by the benchmark, the card points to the right order-theoretic argument: the family of rigid subsets is nonempty and finite, so a minimal member exists.

Mechanism label. theorem selection / case decomposition.

G.3. Off-Rank and Random Distractor Samples

Distractor-1: off-rank distractor, problem 303

Case ID / domain. Distractor-1; complex analysis.

Problem requirement. Compute the same large-arc argument change for $p(z) = z^4 + z^3 + 4z^2 + 3z + 7$.

Retrieved skill. A rank-2 card recommends the argument principle over a closed contour for counting zeros in a sector.

Key answer trace. Top-1 retrieval returns 2π. Rank-2 retrieval returns 0.

Mathematical diagnosis. The off-rank card is broadly related to argument changes, but it changes the task from open-arc asymptotics to a closed-contour zero-counting template. The solver then tracks the wrong path and loses the leading-term arc contribution.

Mechanism label. template overfitting / representation mismatch.

Distractor-2: off-rank distractor, problem 626

Case ID / domain. Distractor-2; analytic geometry.

Problem requirement. Count lattice points strictly inside the bounded conic region; the benchmark answer is 13.

Retrieved skill. A lower-ranked conic card discusses distance-difference properties of hyperbolas and extremal points on conics.

Key answer trace. Top-1 retrieval returns 13. Rank-5 retrieval returns 17.

Mathematical diagnosis. The replacement card keeps the solver in the conic domain but not on the right counting task. The resulting enumeration includes boundary points that should be excluded by the word "inside," increasing the count from 13 to 17.

Mechanism label. scope mismatch / numeric drift.

Distractor-3: off-rank distractor, problem 791

Case ID / domain. Distractor-3; enumerative combinatorics.

Problem requirement. Count rectangle dissections of a staircase of height n with pairwise distinct rectangle areas; the benchmark answer is 2^{n-2}.

Retrieved skill. A lower-ranked grid-counting card recommends a boundary-selection strategy for rectangle arrangements.

Key answer trace. Top-1 retrieval returns 2^{n-2}. Rank-5 retrieval returns 2^{n-1}.

Mathematical diagnosis. The card is topically related to grid rectangles, but it uses a shifted staircase convention. That adds one binary choice too many, producing an answer off by a factor of two.

Mechanism label. convention shift / numeric drift.

Distractor-4: random distractor, problem 964

Case ID / domain. Distractor-4; multivariable calculus.

Problem requirement. Evaluate $\iint_R (x+y)^9(x-y)^9\,dx\,dy$ over $0 \le x \le 1/2$ and $x \le y \le 1-x$; the benchmark answer is $-1/400$.

Retrieved skill. A random card discusses Sard's theorem and dimensional constraints for smooth maps, which is unrelated to the integral.

Key answer trace. Top-1 retrieval and Direct return $-1/400$. Random-skill retrieval returns $+1/400$.

Mathematical diagnosis. The random card provides no useful structure for the change of variables. The solver still attempts a transformation but flips the sign by treating $x-y$ as nonnegative on a region where $y \ge x$.

Mechanism label. numeric/symbolic drift.

Distractor-5: off-rank distractor, problem 1972

Case ID / domain. Distractor-5; algebra / complex roots.

Problem requirement. Find all real a such that four complex roots of a quartic form the vertices of a parallelogram; the benchmark answer is 3.

Retrieved skill. A rank-5 card suggests substitutions for reciprocal power sums such as $z^n + z^{-n}$.

Key answer trace. Top-1 retrieval returns 3. Rank-5 retrieval returns $1, 3$.

Mathematical diagnosis. The off-rank card points to an algebraic manipulation pattern that is only loosely related to the root-geometry condition. It introduces an extra parameter value that does not satisfy the parallelogram constraint.

Mechanism label. template overfitting / symbolic drift.