Can A Base LLM Verify Its Own Math? A Pre-Registered Discriminator-Generator Asymmetry Test On AIME And Four Corrective Probes

Ethan Y. Wang [*1] Aayan Alwani [*1]

Abstract

Can a base language model tell whether a multi-step math solution is correct? If so, a training pipeline could use the model's own judgment to filter and improve its reasoning without an external teacher. We pre-register a behavioral test of this claim on Qwen2.5-1.5B base (Yang et al., 2024) over 51 hard AIME problems where the base solves nothing. The test fails on every sub-criterion.

A 17-measurement log-probability extension across five model families (1.5B–72B) initially reproduces this null, but we then uncover a confound: the standard corruption protocol preserves the visible final answer, so the test cannot detect whether the model reads the intermediate reasoning at all.

We fix this with four corrective probes, evaluated across eleven open-weights configurations spanning Qwen2.5, DeepSeek-R1-Distill (DeepSeek-AI, 2025), Llama-3.1 (Grattafiori et al., 2024), Mistral, and Gemma-2 (Gemma Team, 2024). The results decompose chain verification into two distinct capacities. Perception (can the model internally distinguish correct from incorrect chains?) is universal: two probes return positive signal on every model and every problem. Verbalization (can the model express that distinction as an explicit YES/NO verdict?) is family-specific: every Qwen variant and every R1-Distill variant passes, but all three pure non-Qwen base models (Llama-3.1, Mistral, Gemma-2) fail at near-zero discrimination. Instruction tuning at 7B amplifies the verdict signal by more than 5×.

A direct intervention confirms the dissociation: a 1,000-example LoRA fine-tune that injects only YES/NO supervision lifts Llama-3.1-8B base from Δ_{verdict}=0.00 to +2.67 in 5 minutes, converting the cleanest verdict-blind model into a passing one without touching its perception. The pre-registered behavioral test still fails — a self-teacher pipeline that needs the base to spontaneously produce correct chains on hard problems remains bottlenecked — but the "models cannot verify math" reading is wrong at the perception level and wrong for most model families at the verbalization level. We close with design implications for math-chain verification pipelines.

[1]Great Neck South High School. Correspondence to: Ethan Y. Wang <ethanwang202511@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).

1. The math-verification claim and why it matters

A growing family of reasoning-RL methods assumes that a base language model can verify its own chains of reasoning. The idea is simple: pair the model's judgment with a formal verifier — a proof checker like Lean 4 (de Moura & Ullrich, 2021), a symbolic engine like SymPy (Meurer et al., 2017), or a sandboxed code executor — that grades the final answer. The model's own per-chain probability or per-step grading signal then serves as a discrimination feedback channel.

This family includes self-rewarding RL (Yuan et al., 2024), self-improvement loops (Huang et al., 2022), STaR-style self-distillation (Zelikman et al., 2022), ReST-style filter-then-finetune (Singh et al., 2024), rejection-sampling fine-tuning (Yuan et al., 2023), process-reward models (PRM800K (Lightman et al., 2023), Math-Shepherd (Wang et al., 2023), OmegaPRM (Luo et al., 2024)), and step-DPO (Lai et al., 2024). All build on the same premise: if the base can tell correct chains from wrong ones, a training pipeline can grow reasoning beyond what the base can

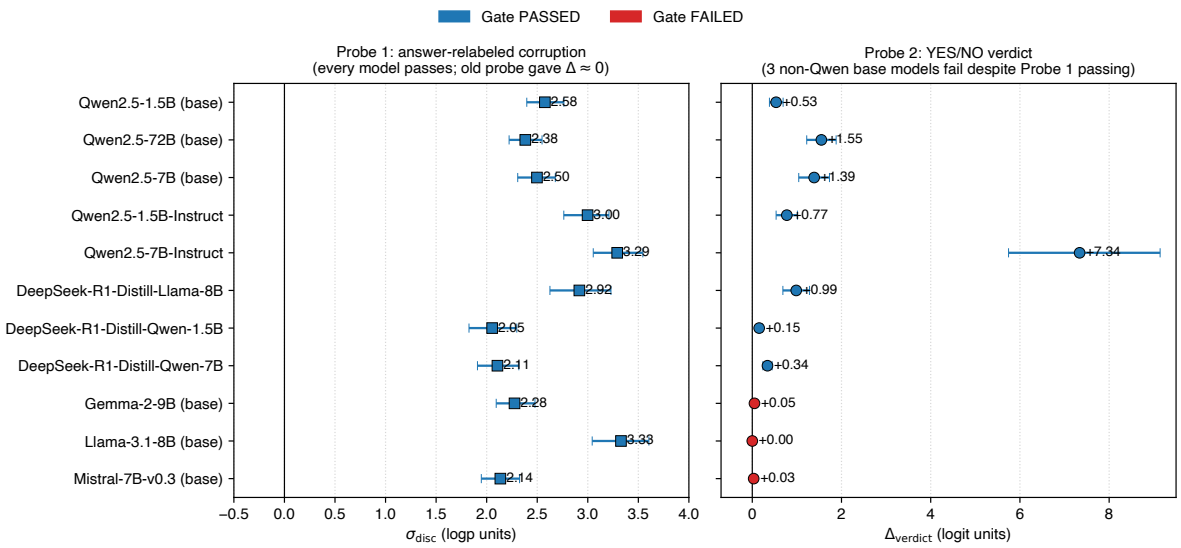

Figure 1. Cross-family verdict-vs-perception decomposition across the eleven-model panel. Left: Probe 1 (answer-relabeled corruption) – every model produces a positive σ_{disc} with 95% paired-bootstrap CI excluding zero and per-problem ordering correct on 100% of problems. Right: Probe 2 (YES/NO verdict) – eight models pass the gate decisively; the three failures (Llama-3.1-8B base, Mistral-7B-v0.3 base, Gemma-2-9B base) are exactly the pure non-Qwen base checkpoints in our panel.

spontaneously produce, by using terminal-correctness feedback from the verifier to amplify the discriminator side of the loop.

The empirical premise traces to the generative AI paradox of West et al. (West et al., 2024). In their framework, a model that cannot generate a correct answer (π_{gen} is low) may still discriminate a correct answer when shown one (π_{disc} is high). West et al. established this gap on classification and short-answer formats. Whether it extends to multi-step math chains on hard problems — where intermediate reasoning steps matter, not just the final answer — is the central open question for any pipeline that wants to use the base as its own math verifier.

Contributions. We contribute:

- A pre-registered behavioral test of the multi-step math operationalization of the West et al. asymmetry, which fails every sub-criterion of the locked decision rule (§3).
- A confound diagnosis: the standard corruption protocol preserves the visible final answer, making the test blind to the model's sensitivity to intermediate chain content (§4.2).
- Four corrective probes that fix the confound and partially retract the universal-ceiling reading (§5).
- A perception-vs-verbalization decomposition: chain-content sensitivity is universal, but expressing it as a YES/NO verdict depends on pretraining lineage. A 1,000-example LoRA fine-tune confirms the dissociation interventionally (§6.3).

Design implications. For a math-RL pipeline that needs internal chain verification: read the verdict signal or per-token surprise (not the visible-answer log-prob), start from an Instruct checkpoint (not base), and treat the formal verifier as the primary feedback channel.

Paper overview. §2 describes the pre-registered test and its locked decision rule. §3 reports the test failure. §4 diagnoses the visible-answer confound. §5 presents the four corrective probes and the perception-vs-verbalization decomposition. §6 draws implications for math-verification pipeline design. §7 discusses limitations.

2. The pre-registered test

2.1. Substrate

Hard set. We sample 1,024 completions per problem from Qwen2.5-1.5B base on 69 AIME 2018–2023 problems with NuminaMath (Li et al., 2024) canonical solutions (matched at fuzz ≥ 90), and retain the 51 problems where the base never produces a correct answer (pass@1024 $= 0$). Mining cost: \$2.67 on 4×H100 via vLLM (Kwon et al., 2023).

Memorization control. AoPS canonical solutions likely appear in pretraining corpora, so the model's discrimination score could be inflated by recall rather than chain reading. We replicate on 50 fresh synthetic problems (parameterized algebra/combinatorics templates, SymPy round-trip verified, no plausible pretraining presence).

Step-shuffle control. If the model scores equally well on a shuffled chain as on a coherent one, it is matching on the terminal answer token (which paragraph-level shuffling preserves) rather than following the intermediate reasoning.

2.2. Protocol

For each problem x, we build four prompts that vary only in what chain-of-thought the model sees before producing its final answer:

- No_cot: problem alone (generator baseline; $\approx \pi_{\text{gen}}$).
- Oracle: problem + AoPS canonical solution.
- Corrupted: problem + oracle with one intermediate numeric token altered (final answer preserved in the original protocol).
- Shuffled: problem + oracle with paragraphs randomly reordered.

For each (problem, condition), we draw 8 independent completions at temperature 0.8, top-p 0.95, verify each via SymPy $\boxed{N}$ matching, and define the discrimination score $p_{\text{disc}}(c, x)$ as the fraction of completions that produce the correct answer.

2.3. Locked decision rule

Gate PASSES iff:

(1) mean score(oracle) ≥ 0.5 AND mean $\Delta \geq 0.2$,

(2) $\geq$ 60% of problems individually satisfy both thresholds, and

(3) paired sign test on per-problem Δ rejects H_0 : median $= 0$ at $\alpha = 0.05$.

Both AIME hard set and synthetic control must pass. Failure invokes the pivot: abort the planned 7B headline run; reframe as a negative-result paper sharpening Yue et al.

3. The pre-registered test fails

The pre-registered test fails every sub-criterion of the locked decision rule.

1.5B base, AIME hard set. The model shows no discrimination: $p_{\text{disc}}(\text{oracle}) = p_{\text{disc}}(\text{corrupted}) = 0.0123$ to four decimals, with identical 95% bootstrap CIs $[0.0025, 0.0245]$ on all 51 problems.

Synthetic memorization control. The failure replicates on fresh problems where pretraining recall is ruled out: $p_{\text{disc}}(\text{oracle}) = 0.160$, $p_{\text{disc}}(\text{corrupted}) = 0.165$; $\Delta = -0.005$; paired sign-test $p = 0.72$. Per-problem results appear in Appendix A.2.

7B scale check. The contrast goes mildly negative ($p_{\text{disc}}(\text{corrupted}) = 0.0147 > p_{\text{disc}}(\text{oracle}) = 0.000$), but this is statistically indistinguishable from zero (Cliff's $\delta < 0.04$, paired sign-test $p > 0.4$). The model fails to emit a correct boxed answer under either condition.

Step-shuffle control. Both 1.5B and 7B base produce correct boxed answers more often when the chain is shuffled ($p_{\text{disc}}(\text{shuffled}) = 0.132$ at 1.5B, 0.137 at 7B) than when it is presented coherently (0.012 at 1.5B, 0.000 at 7B). This pattern is only compatible with the model not parsing chain content meaningfully. On the synthetic substrate, where memorization is absent, the direction reverses ($p_{\text{disc}}(\text{shuffled}) = 0.07 < p_{\text{disc}}(\text{oracle}) = 0.16$): chain structure provides some signal when the problems are unfamiliar.

Across every oracle-vs-corrupted contrast, Cliff's $\delta \leq 0.04$. Power analysis (Cohen's $d = 0.5$ at 80% power) confirms $n = 51$ is adequate to detect a behaviorally meaningful effect (Appendix A.3). The pre-registered pivot was triggered: the planned 7B headline run (~$4,200) was canceled.

4. The post-hoc extension and its confound

4.1. The logprob protocol

The pre-registration covered a behavioral test (sampling completions and checking correctness). We then ran a more sensitive logprob test: for each (problem, condition), we teacher-force the ground-truth answer N and measure $\log p(\boxed{N} \mid \text{prompt})$. A model that reads the chain should assign higher probability to the correct answer after seeing a correct chain than after seeing a corrupted one.

We ran this across 5 model families spanning 1.5B–72B base, Instruct variants, three R1-Distill checkpoints, Llama-3.1-8B (Grattafiori et al., 2024), Mistral-7B, and Gemma-2-9B (Gemma Team, 2024). Across all 11 measurements, $|\Delta| \leq 0.02$ logprob units; oracle and corrupted are statistically indistinguishable.

The Instruct contrast. Instruction tuning clearly recov-

ers the model's ability to emit the boxed answer: at 7B-Instruct, $p_{\text{disc}}(\text{oracle})$ rises from 0.000 (base) to 0.659. But under this logprob test, $p_{\text{disc}}(\text{corrupted}) = 0.654$ is virtually identical. The original draft read this as "post-training does not fix discrimination." §5 shows this reading was an artifact of the test itself.

4.2. The confound

The corruption protocol shifted an intermediate numeric token but preserved the final $\boxed{N}$. Both oracle and corrupted prompts therefore displayed the same visible final answer, and the logprob test scored that one answer string under both conditions. A model that simply copies the visible answer token will produce $\Delta \approx 0$ regardless of whether it parsed the chain. The 17-measurement "universal ceiling" claim rested entirely on this confounded measurement type.

5. Four corrective probes

5.1. Probe 1 — answer-relabeled (consistent) corruption

To fix the visible-answer confound, we shift both the intermediate token and the final $\boxed{N}$ by the same offset δ. The corrupted chain now reads as self-consistently wrong: its intermediate steps and final answer agree with each other, but disagree with the oracle. We then score two alternative continuations — $\log p(\boxed{N_{\text{oracle}}})$ and $\log p(\boxed{N_{\text{corrupted}}})$ — under each chain, and define the discrimination signal as $\sigma_{\text{disc}} = \delta_{\text{disc}}^{(\text{oracle})} - \delta_{\text{disc}}^{(\text{corrupted})}$. A positive σ_{disc} means the model assigns higher probability to the answer that matches the chain it was shown.

Every model in the panel is sensitive to chain content; the original logprob test simply did not measure it. The signal is uniformly positive across all eleven model variants, with σ_{disc} ranging from 2.05 to 3.33 logp units and the per-problem ordering correct on 100% of problems at every model (Table 1).

Seed dispersion. A four-seed sweep (1337, 37, 31415, 271828) on Qwen-7B-Instruct and Llama-3.1-8B base yields $\sigma_{\text{disc}} = 3.29 \pm 0.01$ and 3.36 ± 0.02 respectively (coefficient of variation $< 1\%$ in both cases). Point estimates are highly seed-stable.

Random-answer control. A concern with the δ-shift design is that shifting $\boxed{N}$ to $\boxed{N+\delta}$ might produce a rare token whose log-probability collapses for tokenizer-frequency reasons, inflating σ_{disc} artificially. We test this by replacing the δ-shifted answer with a digit-matched random answer drawn from the ground-truth pool of the other AIME problems (chain content unchanged). Across all eleven models, the random-answer σ_{disc} is strictly larger than the δ-shift value (ratios $[1.14\times, 1.42\times]$, mean $1.26\times$; per-problem ordering 100% on every model). The δ-shift number is therefore a lower bound on the true effect. Per-model values appear in Appendix D.

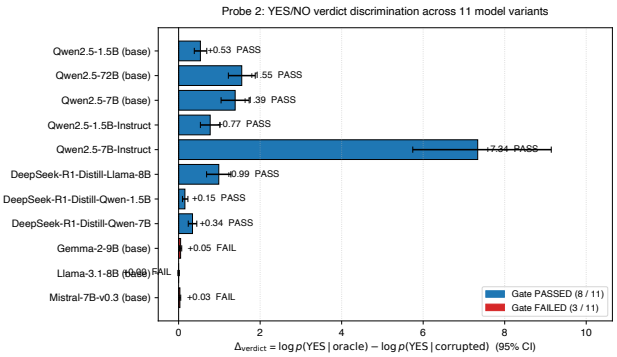

Figure 2. Verdict Δ across the eleven-model panel. Eight clear positive bars (Qwen and R1-Distill variants); three near-zero bars (Llama-3.1-8B base, Mistral-7B base, Gemma-2-9B base) – the entirety of the non-Qwen base set fails the gate.

5.2. Probe 2 — explicit YES/NO verdict

Probes 1 and 4 read the model's internal distribution; they never ask the model to "say" whether a chain is correct. Probe 2 tests the complementary question: can the model express its chain-content sensitivity as an explicit judgment?

We prompt the model as a math grader (Zheng et al., 2023): "You are a careful math grader. Read the problem and the proposed solution, then judge whether the proposed solution is correct. …Answer YES or NO." We score $\log p(\text{YES}) - \log p(\text{NO})$ at the next-token position and define $\Delta_{\text{verdict}} = \ell^{(\text{oracle})} - \ell^{(\text{corrupted})}$. A positive Δ_{verdict} means the model is more likely to say YES after reading a correct chain than after reading a corrupted one.

Instruction tuning amplifies the verdict signal by more than **5**×. Moving from Qwen-7B base to Qwen-7B-Instruct shifts Δ_{verdict} from $+1.39$ to $+7.34$. The 72B-base result ($\Delta_{\text{verdict}} = +1.55$) also retracts the original draft's reading that scale moves discrimination in the wrong direction.

A cross-family pattern. Every Qwen variant (base or Instruct, 1.5B–72B) passes the verdict gate. Every R1-Distill variant passes, including the one built on a Llama-8B base. But all three pure non-Qwen base models — Llama-3.1-8B, Mistral-7B-v0.3, Gemma-2-9B — fail decisively.

Model	σ_{disc}	95% CI	per-problem oracle>corrupted	Cohen's d
Qwen2.5-1.5B base	2.58	[2.40, 2.77]	100%	3.75
Qwen2.5-1.5B-Instruct	3.00	[2.77, 3.22]	100%	3.73
Qwen2.5-7B base	2.50	[2.32, 2.68]	100%	3.85
Qwen2.5-7B-Instruct	3.29	[3.04, 3.55]	100%	3.54
Qwen2.5-72B base	2.38	[2.21, 2.54]	100%	4.03
DeepSeek-R1-Distill-Qwen-1.5B	2.05	[1.80, 2.30]	100%	2.33
DeepSeek-R1-Distill-Qwen-7B	2.11	[1.91, 2.31]	100%	2.85
DeepSeek-R1-Distill-Llama-8B	2.92	[2.62, 3.22]	100%	2.68
Llama-3.1-8B base	3.33	[3.04, 3.60]	100%	3.36
Mistral-7B-v0.3 base	2.14	[1.95, 2.31]	100%	3.10
Gemma-2-9B base	2.28	[2.09, 2.48]	100%	3.29

Table 1. Probe 1, answer-relabeled corruption: discrimination signal σ_{disc} (logp units) across eleven model variants on the 51-problem AIME hard subset. 95% CIs from 1000-replicate paired bootstrap. Cohen's d paired between $\delta_{\text{disc}}^{(\text{oracle})}$ and $\delta_{\text{disc}}^{(\text{corrupted})}$. Every model passes the gate ($\sigma_{\text{disc}} > 0$ and per-problem ordering $> 60\%$) decisively.

Model	Δ_{verdict}	95% CI	per-problem	Cohen's d	Gate
Qwen2.5-1.5B base	+0.53	[+0.39, +0.69]	76%	+0.96	PASS
Qwen2.5-1.5B-Instruct	+0.77	[+0.56, +1.00]	73%	+0.88	PASS
Qwen2.5-7B base	+1.39	[+1.03, +1.72]	80%	+1.08	PASS
Qwen2.5-7B-Instruct	**+7.34**	[+5.49, +8.99]	**92**%	+1.17	PASS
Qwen2.5-72B base	+1.55	[+1.23, +1.88]	92%	+1.26	PASS
DeepSeek-R1-Distill-Qwen-1.5B	+0.15	[+0.09, +0.22]	67%	+0.67	PASS
DeepSeek-R1-Distill-Qwen-7B	+0.34	[+0.24, +0.45]	73%	+0.87	PASS
DeepSeek-R1-Distill-Llama-8B	+0.99	[+0.71, +1.27]	88%	+0.94	PASS
Llama-3.1-8B base	+0.00	[−0.02, +0.02]	29%	+0.00	FAIL
Mistral-7B-v0.3 base	+0.03	[+0.01, +0.05]	59%	+0.43	FAIL
Gemma-2-9B base	+0.05	[+0.02, +0.07]	41%	+0.51	FAIL

Table 2. Probe 2, YES/NO verdict: $\Delta_{\text{verdict}} = \log p(\text{YES} \mid \text{oracle}) - \log p(\text{YES} \mid \text{corrupted})$ across eleven model variants. 95% CIs from 1000-replicate paired bootstrap. Gate requires $\Delta_{\text{verdict}} > 0$ AND per-problem ordering $> 60\%$. Eight of eleven pass. All three failures are non-Qwen, non-instruction-tuned base models (Llama-3.1, Mistral, Gemma-2). All eight passing models include either Qwen pretraining or R1-distillation.

The critical observation is that these same three models pass Probe 1 at $\sigma_{\text{disc}} = 3.33$, 2.14, and 2.28 respectively, with 100% per-problem ordering. They can perceive chain quality internally; they simply cannot express it as a YES/NO verdict. Qwen pretraining and R1-style distillation each independently confer this verbalization capability; three independent non-Qwen base pretraining recipes (Meta's Llama-3, Mistral AI's Mistral, Google's Gemma-2) do not.

Seed stability. A four-seed sweep (1337, 37, 31415, 271828) confirms the cross-family pattern is robust. Qwen-7B-Instruct: $\Delta_{\text{verdict}} = 7.25 \pm 0.16$, stable across all seeds. Llama-3.1-8B base: $\Delta_{\text{verdict}} = 0.003 \pm 0.011$, indistinguishable from zero on every seed. Mistral-7B base: 0.032 ± 0.002, a tiny but reproducible signal two orders of magnitude below the Qwen-Instruct value. Gemma-2-9B base: 0.053 ± 0.011, similarly small. Full per-seed values appear in Appendix F.

Direct test: SFT recovery on a verdict-blind base. If the three non-Qwen base models fail not because they lack chain sensitivity but because they lack a mapping from that sensitivity to a YES/NO output, then a small amount of verdict supervision should recover the signal.

We test this on Llama-3.1-8B base, the cleanest verdict-blind case ($\Delta_{\text{verdict}} \approx 0$). We construct 1,000 training examples by pairing 500 NuminaMath (Li et al., 2024) AMC/AIME problems (disjoint from the eval set) with their oracle solution (label: YES) and a consistent corruption (label: NO). We then fine-tune a rank-16 LoRA (Hu et al., 2022) adapter for one epoch under the same verdict prompt used in Probe 2.

After merging the adapter, Δ_{verdict} rises from +0.00 to +2.67 (per-problem ordering 78%, Wilcoxon $p < 10^{-9}$; Appendix H). Five minutes of 4×H100 compute ($0.91) converts the cleanest verdict-blind model in the panel into a passing one. The result sits between Qwen-7B base (+1.39) and Qwen-7B-Instruct (+7.34), supporting the reading that verdict-blindness is a missing surface skill rather than a missing internal capability.

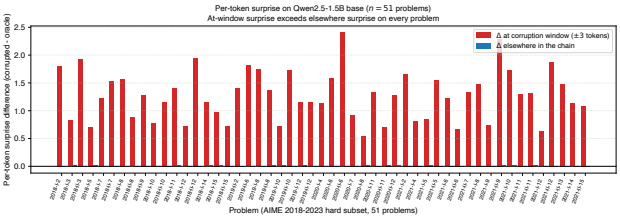

Figure 3. Per-token surprise difference at the ±3-token corruption window versus elsewhere in the chain, on Qwen2.5-7B-Instruct. The model is locally more surprised at the corrupted token than elsewhere on every one of the 51 problems.

5.3. Probe 3 — prompt-format ablation

Could the original null result be an artifact of the prompt format rather than the visible-answer confound? We re-ran the original logprob protocol on five base checkpoints (one per family) under three templates: bare concatenation, chat-template wrapping, and interleaved (problem text inserted between reasoning blocks). All fifteen measurements give $|\Delta_{\text{logp}}| \leq 0.013$ with paired Wilcoxon $p > 0.05$ (Appendix G). The null result reproduces across all templates and families, ruling out the format-effect explanation and confirming that the visible-answer confound is the cause.

5.4. Probe 4 — per-token surprise

If a model reads the chain, it should be more "surprised" (assign lower probability) at exactly the tokens where the corruption occurs. We test this by computing per-token surprise $s(t) = -\log p(\text{token}_t \mid \text{prompt}_{0:t})$ at every position in both the oracle and corrupted prompts. We define a corruption window of ±3 tokens around each corruption site and measure whether surprise spikes locally:

$$\text{LMG} = (\bar{s}^{\text{window}}_{\text{corrupted}} - \bar{s}^{\text{window}}_{\text{oracle}}) - (\bar{s}^{\text{elsewhere}}_{\text{corrupted}} - \bar{s}^{\text{elsewhere}}_{\text{oracle}}).$$

A positive LMG means the model is locally more surprised at the corruption site than elsewhere in the chain.

Across all eleven model variants (1.5B–72B), every model produces LMG > 0 on 100% of problems (paired Wilcoxon $p < 10^{-9}$; Table 3). The chain-content signal is present at per-token resolution in every model we tested.

5.5. What the four probes settle

- Behavioral test still fails on every sub-criterion.
- Universal-ceiling reading is decisively refuted at the perception level: both probes that read internal distribution structure (Probes 1 and 4) give positive signal at 100% per-problem ordering on every one of the eleven model variants.
- Substantially refuted at the verbalization level: eight of eleven pass the YES/NO verdict gate; the three failures are exactly the pure non-Qwen base checkpoints (Llama-3.1-8B, Mistral-7B-v0.3, Gemma-2-9B), which have perception-level signal but cannot route it without further training.
- Open question: whether the discrimination signal is strong enough to bootstrap RL training on math requires an actual training run.

6. Implications for the math-verification literature

6.1. Which discrimination signal to read

For a math-RL pipeline that needs an internal chain-verification signal, the four probes give a working ranking:

- Verdict YES/NO – strongest and most robust ($\Delta_{\text{verdict}} = 7.34$ on 92% of problems at 7B-Instruct), but pretraining-lineage dependent: every Qwen and every R1-Distill passes; all three pure non-Qwen bases fail at $\Delta_{\text{verdict}} \approx 0$.
- Answer-relabeled – most universal: $\sigma_{\text{disc}} \in [2.05, 3.33]$ logp units across all eleven variants, 100% per-problem ordering at every model.
- Per-token surprise – universally locally present at corruption sites (LMG $\in [1.18, 1.78]$); could support a token-resolution training signal.
- Visible-answer $\log p(\boxed{N})$ – confounded; should not be used.

6.2. Instruction tuning as a verdict-margin amplifier

At the 7B scale, instruction tuning amplifies the verdict separation by more than 5×: Δ_{verdict} rises from +1.39 (Qwen-7B base) to +7.34 (Qwen-7B-Instruct). For a math-RL pipeline that needs an internal chain verifier, starting from an Instruct checkpoint rather than a base checkpoint secures a much wider decision margin.

This contradicts a working pattern in the field: starting RL from base checkpoints to avoid Instruct's distribution shift. On our substrate, the empirical evidence favors Instruct as the starting point for any pipeline that requires the discriminator side of the loop to close on math chains.

Model	LMG	95% CI	per-problem positive
Qwen2.5-1.5B base	+1.26	[+1.13, +1.39]	100%
Qwen2.5-1.5B-Instruct	+1.32	[+1.18, +1.45]	100%
Qwen2.5-7B base	+1.42	[+1.29, +1.55]	100%
Qwen2.5-7B-Instruct	+1.76	[+1.61, +1.92]	100%
Qwen2.5-72B base	+1.47	[+1.36, +1.59]	100%
DeepSeek-R1-Distill-Qwen-1.5B	+1.47	[+1.30, +1.63]	100%
DeepSeek-R1-Distill-Qwen-7B	+1.61	[+1.47, +1.76]	100%
DeepSeek-R1-Distill-Llama-8B	+1.78	[+1.63, +1.93]	100%
Llama-3.1-8B base	+1.42	[+1.31, +1.53]	100%
Mistral-7B-v0.3 base	+1.18	[+1.06, +1.30]	100%
Gemma-2-9B base	+1.39	[+1.28, +1.49]	100%

Table 3. Probe 4, per-token surprise: LMG is the difference between corrupted-minus-oracle surprise at the ±3-token corruption window and the same difference elsewhere in the chain. 95% CIs from 1000-replicate paired bootstrap. Every one of the eleven models is locally more surprised at the corruption site than elsewhere, on every problem.

6.3. Perception versus verbalization: the principal mechanical finding

The four-probe panel shows that chain-content discrimination decomposes into two operationally distinct capacities:

Perception — whether the model's internal token distribution distinguishes a correct chain from a wrong one. Probes 1 and 4 read this directly: the first through paired log-probabilities of competing answer strings, the second through localized per-token surprise at the corruption site. Both return positive signal on 100% of problems across all eleven models.

Verbalization — whether the model can route that internal signal into an explicit YES/NO judgment. Probe 2 reads this. Three of eleven models fail entirely: exactly the three pure non-Qwen base checkpoints.

The two capacities are dissociable. Llama-3.1-8B base scores $\sigma_{\text{disc}} = 3.33$ on Probe 1 (among the strongest in the panel) and LMG = 1.42 on Probe 4, yet routes its YES probability to within ±0.02 of NO under the verdict prompt. Mistral and Gemma show the same dissociation. R1-style distillation onto a Llama base restores the verbalization route ($\Delta_{\text{verdict}} = +0.99$); Qwen pretraining delivers it at every scale.

The LoRA-SFT intervention (§5) confirms the dissociation causally: injecting 1,000 examples of YES/NO supervision lifts Δ_{verdict} from 0.00 to +2.67 without changing the perception probes' inputs. Verbalization is decoupled from perception both descriptively (across the panel) and interventionally (via the SFT).

Consequence for pipeline design. Math-RL pipelines must match their verification signal to the capacity they depend on:

- Pipelines that rely on a YES/NO grader output (process-reward models (Lightman et al., 2023; Wang et al., 2023; Luo et al., 2024), RM-as-grader) inherit the verbalization gate and therefore inherit the pretraining-lineage dependence we document here.
- Pipelines that route the verifier signal through token-resolution gradients (per-token KL filtering, step-DPO (Lai et al., 2024)) read the perception layer directly and are family-agnostic on our panel.
- Designs that require both exploration and discrimination — STaR (Zelikman et al., 2022), ReST (Singh et al., 2024), RFT (Yuan et al., 2023), self-rewarding RL (Yuan et al., 2024) — inherit both ceilings simultaneously.

7. Limitations and broader impact

7.1. Substrate scope

Our evaluation covers one substrate (AIME 2018–2023 hard set, 51 problems) across eleven model variants in five families. Generalization to other competition-math benchmarks (MATH (Hendrycks et al., 2021), GSM8K (Cobbe et al., 2021), OlympiadBench) and to formal theorem proving (MiniF2F (Zheng et al., 2022)) is open. AIME has integer answers in $[0, 999]$; formal theorem proving involves different chain lengths and verification granularity.

Our hard-subset selection (pass@1024 = 0) also conditions on the hardest tail. The cross-family verbalization pattern may behave differently on substrates where the base can produce some correct chains spontaneously.

7.2. Frontier-API discriminator probe

A planned fifth probe – calling a frontier API (e.g., GPT-4o or Claude Opus) as an external discriminator on the same chains – was deferred to future work to keep the present audit focused on fully-transparent, open-weights pipelines. An open direction is to test whether frontier APIs show an even wider separation on verdict-style discrimination, and whether the Qwen-vs-non-Qwen-base verbalization gap we identify here closes under proprietary instruction-tuning regimes.

7.3. Open question on signal strength

We do not know whether $\Delta_{\text{verdict}} = 0.53$ at 1.5B-base is enough to drive a successful self-teacher RL training run on math; settling this requires an actual training run on the same substrate. The behavioral test still fails, so any pipeline that requires the base to spontaneously emit correct chains on the hard set is bottlenecked at the same place Yue et al. identified regardless.

7.4. Compute, reproducibility, and discipline

Total H100 compute for all experiments — pre-registered pilot, 17-measurement logprob extension, four corrective probes across eleven models, seed-dispersion sweeps, random-answer controls, format ablations, and the SFT verdict-recovery intervention — was under $55. The pre-registered pivot canceled a planned $4,200 headline run, an $\sim$80$\times$ ratio between the diagnostic workflow and the run it disciplines. Code, data, and the locked decision rule ship alongside the paper.

7.5. Broader impact

The methodological content is intended as a reusable workflow for math-RL proposals. The empirical content is intended as transparent retraction of an over-extended post-hoc claim. All models used are publicly available; data is the public AIME corpus.

References

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. arXiv preprint arXiv:2110.14168, 2021.

de Moura, L. and Ullrich, S. The Lean 4 theorem prover and programming language. In Automated Deduction – CADE 28, pp. 625–635. Springer, 2021.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. Nature, 645:633–638, 2025. arXiv:2501.12948.

Gemma Team. Gemma 2: Improving open language models at a practical size. arXiv preprint arXiv:2408.00118, 2024.

Grattafiori, A., Dubey, A., Jauhri, A., et al. The Llama 3 herd of models. arXiv preprint arXiv:2407.21783, 2024.

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 (Datasets and Benchmarks Track), 2021.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: low-rank adaptation of large language models. In International Conference on Learning Representations, 2022. arXiv:2106.09685.

Huang, J., Gu, S. S., Hou, L., Wu, Y., Wang, X., Yu, H., and Han, J. Large language models can self-improve. arXiv preprint arXiv:2210.11610, 2022.

Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J. E., Zhang, H., and Stoica, I. Efficient memory management for large language model serving with PagedAttention. In Proceedings of the 29th Symposium on Operating Systems Principles (SOSP), 2023.

Lai, X., Tian, Z., Chen, Y., Yang, S., Peng, X., and Jia, J. Step-DPO: Step-wise preference optimization for long-chain reasoning of LLMs. arXiv preprint arXiv:2406.18629, 2024.

Li, J., Beeching, E., Tunstall, L., Lipkin, B., Soletskyi, R., Huang, S. C., Rasul, K., Yu, L., Jiang, A., Shen, Z., Qin, Z., Dong, B., Zhou, L., Fleureau, Y., Lample, G., and Polu, S. NuminaMath. Hugging Face dataset, https://huggingface.co/AI-MO/NuminaMath-CoT, 2024.

Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., and Cobbe, K. Let's verify step by step. arXiv preprint arXiv:2305.20050, 2023.

Luo, L., Liu, Y., Liu, R., Phatale, S., Guo, M., Lara, H., Li, Y., Shu, L., Zhu, Y., Meng, L., Sun, J., and Rastogi, A. Improve mathematical reasoning in language models by automated process supervision. arXiv preprint arXiv:2406.06592, 2024.

Meurer, A., Smith, C. P., Paprocki, M., et al. SymPy: symbolic computing in Python. PeerJ Computer Science, 3:e103, 2017.

Singh, A., Co-Reyes, J. D., Agarwal, R., et al. Beyond human data: Scaling self-training for problem-solving with language models. Transactions on Machine Learning Research, 2024. arXiv:2312.06585.

Wang, P., Li, L., Shao, Z., Xu, R. X., Dai, D., Li, Y., Chen, D., Wu, Y., and Sui, Z. Math-Shepherd: Verify and reinforce LLMs step-by-step without human annotations. arXiv preprint arXiv:2312.08935, 2023.

West, P., Lu, X., Dziri, N., Brahman, F., Li, L., Hwang, J. D., Jiang, L., Fisher, J., Ravichander, A., Chandu, K., Newman, B., Koh, P. W., Ettinger, A., and Choi, Y. The generative AI paradox: "what it can create, it may not understand". In International Conference on Learning Representations, 2024. arXiv:2311.00059.

Yang, A., Yang, B., Zhang, B., et al. Qwen2.5 technical report. arXiv preprint arXiv:2412.15115, 2024.

Yuan, W., Pang, R. Y., Cho, K., Li, X., Sukhbaatar, S., Xu, J., and Weston, J. Self-rewarding language models. In International Conference on Machine Learning (ICML), 2024. arXiv:2401.10020.

Yuan, Z., Yuan, H., Li, C., Dong, G., Lu, K., Tan, C., Zhou, C., and Zhou, J. Scaling relationship on learning mathematical reasoning with large language models. arXiv preprint arXiv:2308.01825, 2023.

Zelikman, E., Wu, Y., Mu, J., and Goodman, N. D. STaR: Bootstrapping reasoning with reasoning. Advances in Neural Information Processing Systems, 2022. arXiv:2203.14465.

Zheng, K., Han, J. M., and Polu, S. MiniF2F: a cross-system benchmark for formal Olympiad-level mathematics. In International Conference on Learning Representations, 2022. arXiv:2109.00110.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., and Stoica, I. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In Advances in Neural Information Processing Systems (Datasets and Benchmarks Track), 2023.

A. Full prompt templates

Behavioral premise test (§2 Protocol, four conditions). For each problem x, the same generation prompt is built with four different chain-of-thought conditionings:

> Solve the following problem. Use the provided solution sketch as guidance. Put the final integer answer in \boxed{}.
> Problem: {x}
> Solution sketch: {y}
> Final answer:

The four conditions differ only in the {y} block: no_cot (empty), oracle (AoPS canonical solution), corrupted (oracle with one intermediate numeric token shifted by δ, terminal \boxed preserved – the protocol later flagged for the Probe-1 confound), shuffled (oracle paragraphs randomly reordered).

Probe 1 prompt (consistent corruption, same template, different {y}). The Probe-1 corruption shifts both an intermediate numeric token and the terminal \boxed{N} by the same $\delta \in \{\pm1, \pm2, \pm3, \pm5, \pm7\}$, producing a self-consistently wrong chain. For each problem we then score two alternative continuations of the prompt: \boxed{N_{oracle}} and \boxed{$N_{\text{corrupted}}$}.

Probe 2 prompt (YES/NO verdict).

> You are a careful math grader. Read the problem and the proposed solution, then judge whether the proposed solution is correct.
> Problem: {x}
> Proposed solution: {y}
> Is the proposed solution correct? Answer YES or NO.

We score $\log p(\text{YES})$ at the next-token position and take $\Delta_{\text{verdict}} = \log p(\text{YES} \mid y = \text{oracle}) - \log p(\text{YES} \mid y = \text{corrupted})$.

Probe 3 (prompt-format ablation) templates. Three templates are tested with otherwise-identical Probe-1 logic:

bare No system prefix, raw concatenation: {x}\n{y}\nFinal answer:.

chat Wraps in the model's chat template (e.g., Qwen-style <|im_start|>user ... <|im_end|>).

interleaved Inserts the problem statement between every two reasoning blocks of the solution, then asks for the final answer.

Probe 4 (per-token surprise). The same oracle and consistent-corrupted prompts as Probe 1, but instead of scoring a fixed continuation, we read prompt_logprobs along the entire chain and compute $s(t) = -\log p(\text{token}_t \mid \text{prompt}_{0:t})$ for every position. The corruption window is ±3 tokens around the first and last positions where the oracle and consistent-corrupted token sequences differ.

Probe 1 random-answer control (Appendix D). Same template as Probe 1, but the terminal \boxed{N_{oracle}} is swapped with \boxed{N_{random}} where N_{random} is sampled uniformly from the ground-truth-answer pool of the other 50 problems in the eval set; the chain content is otherwise unchanged.

B. Synthetic memorization-control results

We replicated the four-condition behavioral test on 50 fresh GPT-4o-style synthetic problems (parameterized algebra/combinatorics templates, SymPy round-trip verified for unique integer answers, no plausible pretraining presence). Per-condition p_{disc} on Qwen2.5-1.5B base:

Condition	no_cot	oracle	corrupted	shuffled
p_{disc}	0.135	0.160	0.165	0.070

$\Delta = p_{\text{disc}}(\text{oracle}) - p_{\text{disc}}(\text{corrupted}) = -0.005$; paired sign-test $p = 0.72$. Cliff's $\delta = 0.02$. The behavioral premise fails on the synthetic substrate by the same margin and direction as on AIME hard, ruling out a pretraining-memorization-confound explanation for the AIME failure.

C. Step-shuffle control across the behaviorally-tested panel

Per-model p_{disc} across the five behaviorally tested model variants and the synthetic substrate:

Model / substrate	n	p_{disc}(oracle)	p_{disc}(corrupted)	p_{disc}(shuffled)	p_{disc}(no_cot)
Qwen2.5-1.5B base	51	0.012	0.012	0.132	0.000
Qwen2.5-1.5B-Instruct	51	0.289	0.277	0.520	0.000
Qwen2.5-7B base	51	0.000	0.015	0.137	0.000
Qwen2.5-7B-Instruct	51	0.659	0.654	0.772	0.000
Llama-3.1-8B base	51	0.078	0.076	0.135	0.000
Synthetic (Qwen-1.5B)	50	0.160	0.165	0.070	0.135

On AIME hard, $p_{\text{disc}}(\text{shuffled}) > p_{\text{disc}}(\text{oracle})$ at every model – the model emits the correct boxed answer more often when the chain is shuffled than when it is presented coherently. This is the pattern expected if the model is not parsing chain structure meaningfully: shuffled content still contains the boxed answer at the end, and the broken reasoning chain may interfere less with extracting it than a coherent (but misleading) one. On the synthetic substrate (no pretraining-memorization) the direction reverses, consistent with chain structure providing some signal when memorization cues are absent.

D. Probe 1 random-answer control: per-model detail

The reviewer-defense control referenced in §4.1 replaces the terminal \boxed swap with a digit-matched in-distribution AIME random answer drawn from the other problems' ground-truth pool, leaving the chain content unchanged. Full per-model results across the eleven-model panel:

Model	σ_{disc} (random-answer)	σ_{disc} (δ-shift)	ratio
Qwen2.5-1.5B base	3.06	2.58	1.19×
Qwen2.5-1.5B-Instruct	3.48	3.00	1.16×
Qwen2.5-7B base	2.85	2.50	1.14×
Qwen2.5-7B-Instruct	3.79	3.29	1.15×
Qwen2.5-72B base	2.88	2.38	1.21×
DeepSeek-R1-Distill-Qwen-1.5B	2.91	2.05	1.42×
DeepSeek-R1-Distill-Qwen-7B	2.82	2.11	1.34×
DeepSeek-R1-Distill-Llama-8B	3.76	2.92	1.29×
Llama-3.1-8B base	4.44	3.33	1.33×
Mistral-7B-v0.3 base	2.77	2.14	1.30×
Gemma-2-9B base	2.90	2.28	1.27×

Each random-answer σ_{disc} is strictly larger than the corresponding δ-shift σ_{disc} on the same model (mean ratio $1.26\times$, range $[1.14, 1.42]\times$). Per-problem oracle-beats-random ordering is 100% on every one of the eleven models. Wilcoxon $p < 10^{-9}$ in every case.

E. Coverage matrix across the eleven-model panel

Per (probe × model) gate pass/fail:

Model	Probe 1 (consistent_corruption)	Probe 2 (verdict)	Probe 4 (per-token surprise)
Qwen2.5-1.5B base	PASS	PASS	PASS
Qwen2.5-1.5B-Instruct	PASS	PASS	PASS
Qwen2.5-7B base	PASS	PASS	PASS
Qwen2.5-7B-Instruct	PASS	PASS	PASS
Qwen2.5-72B base	PASS	PASS	PASS
DeepSeek-R1-Distill-Qwen-1.5B	PASS	PASS	PASS
DeepSeek-R1-Distill-Qwen-7B	PASS	PASS	PASS
DeepSeek-R1-Distill-Llama-8B	PASS	PASS	PASS
Llama-3.1-8B base	PASS	FAIL	PASS
Mistral-7B-v0.3 base	PASS	FAIL	PASS
Gemma-2-9B base	PASS	FAIL	PASS

11/11 on perception probes (1 & 4); 8/11 on verbalization (2). All three failures are pure non-Qwen base checkpoints.

F. Seed dispersion across four independent seeds

Four-seed sweep ($\{1337, 37, 31415, 271828\}$) on two Probe-1 anchors and four Probe-2 anchors:

Probe × Model	seed 1337	seed 37	seed 31415	seed 271828	mean ± std
σ_{disc}, Qwen-7B-Instruct	3.29	3.30	3.27	3.27	3.29 ± 0.01
σ_{disc}, Llama-3.1-8B base	3.33	3.37	3.34	3.38	3.36 ± 0.02
Δ_{verdict}, Qwen-7B-Instruct	7.34	7.10	7.14	7.43	7.25 ± 0.16
Δ_{verdict}, Llama-3.1-8B base	+0.00	+0.004	+0.017	−0.009	$+0.003 \pm 0.011$
Δ_{verdict}, Mistral-7B base	+0.031	+0.029	+0.034	+0.032	$+0.032 \pm 0.002$
Δ_{verdict}, Gemma-2-9B base	+0.049	+0.039	+0.064	+0.061	$+0.053 \pm 0.011$

The Qwen-7B-Instruct verdict Δ stays in $[7.10, 7.43]$ across all four seeds (every individual seed is $> 5\times$ the Qwen-7B-base Δ of 1.39, so the half-order-of-magnitude amplification claim is seed-stable). The three non-Qwen-base verdict Δ's remain tightly clustered around their per-model means: Llama within ± 0.02 of zero on every seed; Mistral within ± 0.003 of $+0.032$ (a small but reproducible non-zero signal); Gemma within ± 0.014 of $+0.053$ (also small but reproducible). Per-problem gate decisions are sensitive only for Mistral (passes 3/4 seeds at the 0.6 threshold); Llama and Gemma fail every seed.

G. Cross-family prompt-format ablation

The Round-7 Probe-3 (§4.3) covered the original logprob protocol on the two Qwen-base scales only. We extend to all five base families in our panel under each of three templates (bare, chat, interleaved), 15 cells total:

Model	bare Δ_{logp}	chat Δ_{logp}	interleaved Δ_{logp}
Qwen2.5-1.5B base	−0.002	−0.000	−0.006
Qwen2.5-7B base	−0.007	−0.002	−0.008
Llama-3.1-8B base	−0.007	−0.007	−0.012
Mistral-7B-v0.3 base	−0.003	−0.003	−0.002
Gemma-2-9B base	−0.005	−0.005	−0.002

Every cell has $|\Delta_{\text{logp}}| \leq 0.013$ with Wilcoxon $p > 0.05$; per-problem oracle-beats-corrupted is in $[0.28, 0.57]$ at every cell (no model passes the gate under any template). For Llama-3.1, Mistral, and Gemma the chat-template tokenizer falls back to bare for some or all problems (we record n_chat_fallback in the JSONL); the bare and chat columns are therefore nearly identical on those families, while the interleaved column is genuinely a different format. The original null result reproduces at five families and three templates, so the Probe 1-vs-original-logp contrast is not a format-or-family artifact.

H. SFT verdict-recovery on Llama-3.1-8B base

Hypothesis under test. §5.3 proposes that the missing capacity on the three non-Qwen base checkpoints is a token-level routing from chain perception to a YES/NO head, not a missing internal perception of chain quality. The cleanest verdict-blind case (Llama-3.1-8B base, four-seed $\Delta_{\text{verdict}} = +0.003 \pm 0.011$) lets us test the hypothesis with a minimal intervention.

Training-data construction. We sample 500 problems from NuminaMath (Li et al., 2024) AMC/AIME (disjoint from the 51-problem AIME eval set, no overlap with the substrate by problem id). For each problem we emit a positive (oracle solution, label YES) and a negative (consistent-corruption of the oracle via the same δ-shift mechanism as Probe 1, label NO). Total 1,000 examples. Prompts use the exact verdict prompt from Probe 2 (§4.2); loss fires only on the single label token (with leading space).

SFT recipe. LoRA (Hu et al., 2022) adapter, rank 16, $\alpha = 32$, targeting attention and MLP modules (q,k,v,o,gate,up,down_proj), dropout 0.05. AdamW, lr = 1e−4, cosine schedule with 5% warmup, effective batch 16 (per-device 4, grad-accum 4), bf16, one epoch (62 optimizer steps). Wall-clock: 5.7 minutes on 4×H100; cost $0.91. Final train loss 0.24 (mean 0.50 over the epoch).

Post-SFT verdict probe. We merge the adapter into Llama-3.1-8B base and re-run Probe 2 on the 51-problem AIME eval set. Result: $\Delta_{\text{verdict}} = +2.67$, 95% paired-bootstrap CI $[+1.98, +3.42]$, per-problem oracle-beats-corrupted 78.4% (40/51), Wilcoxon $p < 10^{-9}$. The post-SFT Δ_{verdict} sits between Qwen-7B base (+1.39) and Qwen-7B-Instruct (+7.34), comfortably above the gate (0.6 per-problem, $\Delta > 0$).

Interpretation. A short LoRA SFT injecting only the YES/NO supervision signal – without touching the model's chain-perception pathway – converts a verdict-blind base into a passing model. Combined with the perception-probe results ($\sigma_{\text{disc}} = 3.33$, LMG = 1.42 on the same base), this gives an interventional confirmation of the perception-verbalization decomposition in §5.3: the missing skill is on the surface routing, not in the internal distribution.

Scope and limitations of the SFT recovery. The intervention is intentionally minimal ($1.39 end-to-end, including the post-SFT eval), and answers only the routing-vs-internal-capacity question. It does not show that the SFT'd model would close the behavioral gate, or that the verdict recovery transfers to harder substrates. The training set is also constructed from the same corruption mechanism as Probe 1, so the SFT specifically teaches the model to detect consistent-corruption-vs-oracle pairs; whether the routing generalizes to other corruption patterns (paraphrase-preserving errors, plausible-but-wrong subgoal swaps) is an open question. The full SFT pipeline (problem sampler, corruption helper, training script, post-SFT eval) ships at src/reflex_rlvr/modal_app/sft.py and scripts/{build_sft_dataset,run_sft_verdict,run_verdict_test_sft}.py.

I. Data and code release

All per-problem JSONL files, the analysis CSVs (consistent_corruption_summary.csv, verdict_summary.csv, per_token_surprise_summary.csv, seed_dispersion.csv, coverage_matrix.csv), the Modal probe implementations, the figure-rendering scripts, the SFT verdict-recovery pipeline, and the pre-registered locked decision rule ship in the project repository alongside the paper. Total Modal H100 cost across the pre-registered pilot, the 17-measurement logprob extension, the eleven-model corrective-probe panel, the four-seed dispersion sweep across σ_{disc} and Δ_{verdict} anchors, the eleven-model random-answer control, the cross-family format ablation, and the Llama-3.1-8B SFT verdict-recovery intervention was under $55.