BEYOND CORRECT LABELS: FEEDBACK, EFFICIENCY, AND COVERAGE

Anonymous authors
Paper under double-blind review

ABSTRACT

Mathematical self-evolution requires supervision that improves answer accuracy while preserving the breadth of problems a model can solve. We propose an oracle-based mathematical co-evolution framework that constructs training targets through program execution and screens their agreement with problem statements. To audit the target–grader–policy interface, we define Active Error Exposure (AEE) as the joint probability of an incorrect target and a nonconstant task-reward group, and pair it with replay diagnostics for lost reference distinctions. The first round raises pass@1 from 36.4% to 64.6% on MATH-500 and from 17.0% to 28.5% on a fixed 500-question Olympiad subset, while retaining high pass@256 coverage. Across three recorded R2 pairs, pass@256 gains over an Agent0 reproduction are 2.4–4.2 points on MATH and 9.2–11.0 on Olympiad; native 8K/16K Olympiad checks do not remove the gap. A label-only intervention further shows a training-side result: changing only conflict targets changes delivered group-relative feedback: E conflict groups are mostly all-positive, whereas M groups are mostly all-negative. The downstream effect of that intervention remains unidentified under the available grading versions. In a native 16K Olympiad check, Oracle-based reaches 64.15% pass@32 at 786 tokens per response, versus 56.87% for Agent0 at 9,905 tokens, a descriptive 12.6-fold realized-token contrast. For Agent0, extra decoding is net negative: 276 decisions are rescued but 409 change from positive to negative from its 4K prefix to 16K.

1 INTRODUCTION

Motivated by the long-term prospect of recursive self-improvement (RSI), we ask whether supervision remains usable, rather than merely correctly labeled, as self-evolution rounds accumulate. We study this question in two rounds of mathematical co-evolution. The system must decide which answers deserve reinforcement. Agent0 (Xia et al., 2025) uses executor majority votes; execution-grounded self-play derives candidate answers from programs (Zhao et al., 2026; Zhou et al., 2026). This choice of supervision source shapes both what the model learns and which problems it continues to solve.

Our *oracle-based mathematical co-evolution* framework trains the executor on program-derived targets screened by fresh statement-only implementations; executor feedback guides the next curriculum round. Oracle-based R1 improves single-sample accuracy while retaining high MATH/Olympiad coverage. All three R2 pairs outperform the Agent0 reproduction in 4,096-token pass@256 on MATH and Olympiad, without implying coverage expansion over Base (Appendix N reports a more limited AIME25 retention result).

Evaluating that supervision also requires inspecting the rewards it delivers, not just the final policy. A mathematically correct target can be graded inconsistently across equivalent answers; an incorrect target can reinforce wrong answers or remove distinctions among them. Prior work distinguishes label accuracy from reward accuracy (Zuo et al., 2026), and single-sample success from high-budget coverage (Chen et al., 2026). Group-relative learning depends on which reward distinctions the current policy encounters. Building on wrong-target reachability (Shao et al., 2025) and advantage-collapse diagnostics (He et al., 2026), we define *Active Error Exposure* (AEE) as the joint probability of an incorrect target and a nonconstant task-reward group. AEE identifies active erroneous supervision; paired reference replay complements it by detecting reward distinctions

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

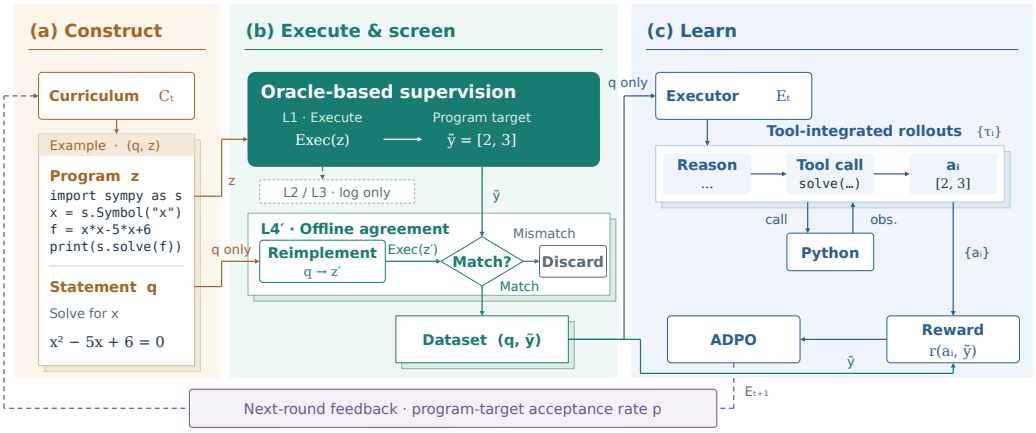

Figure 1: **Oracle-based mathematical co-evolution at round t. (a)** C_t generates statement q and program z (example: roots $[2, 3]$ of $x^2 - 5x + 6 = 0$). **(b)** L1 executes z to obtain $\tilde{y}$; L2/L3 are diagnostics, not gates. Offline L4′ reimplements q as z' without z or $\tilde{y}$, retaining $(q, \tilde{y})$ only when outputs match. Agreement screens consistency, not mathematical correctness. **(c)** E_t receives only q, samples trajectories $\{\tau_i\}$ with answers $\{a_i\}$ using Python observations ("obs."), and receives $r(a_i, \tilde{y})$ for ADPO updates to E_{t+1}. The dashed purple loop returns program-target acceptance rate p as next-round difficulty feedback. Amber, teal and blue arrows denote generated inputs, execution-derived supervision/screening, and executor interactions/learning, respectively.

lost under the training target. Saved-trajectory replay and a same-question target-substitution probe distinguish these cases from equivalent-answer grading failures (Sections 3 and 4.5).

2 Oracle-Based Mathematical Co-Evolution

Figure 1 summarizes how program-derived targets connect curriculum construction, offline screening and executor learning across rounds.

Majority-supervised reference. For a generated question q, an executor samples answers and selects the largest answer-equivalence cluster as its pseudo-label $\tilde{y}$, following the Agent0-style largest-cluster rule. This is a numerical plurality rule: the selected cluster need not contain more than half of the samples, and that is part of the supervision design rather than an absolute-majority requirement. The cluster frequency s measures self-consistency. Agent0 uses this signal for curriculum difficulty and ambiguity-aware policy optimization (ADPO), together with diversity and tool-use incentives (Xia et al., 2025).

Constructing an execution oracle. The curriculum generates program z and statement q, with target $\tilde{y} = \mathrm{Exec}(z)$. The recorded R1/R2-compatible setting uses L1 execution as a gate; L2 repeat-output consistency and L3 answer-leakage checks are diagnostics. Offline L4′ independently implements q and retains $(q, \tilde{y})$ when outputs agree. The executor trains with ADPO without seeing z. We separately audit program execution, statement fidelity and rollout rewards. *Oracle-based* names the supervision design; *execution-grounded* specifies its implementation.

Answer source and difficulty signal are separate choices. The oracle-based curriculum uses program-target acceptance rate p, rather than executor self-consistency s, for difficulty feedback. Appendix D probes this choice and L4′.

Supervision and construction channels. Both agents derive from Qwen3-4B-Base; a stronger model reviews diagnostics, not main training answers (Section 4.5). Construction, filtering, difficulty feedback and answer source change together. Online agreement compares executor votes with program answers; offline agreement uses fresh implementations. ADPO and principal optimization

settings are retained; curriculum outputs supply questions, targets and difficulty metadata. Appendix A records configurations, the resume exception and related executable-supervision designs. L4′ screens targets offline, not executor trajectories; Agent0 is our quantitative reference. The *supervision channel* changes targets/rewards, while the *construction channel* changes tasks/behaviors (Sections 3 and 4.4).

3 Active Error Exposure and Paired Replay

3.1 From verifier quality to policy-dependent exposure

In the framework of Section 2, targets become supervision only through their rewards on executor rollouts. We characterize this interface through target correctness, reward-group activity, and the distinctions changed or lost under reference substitution. This diagnostic framework complements the policy-performance comparisons in Section 4. TTRL (Zuo et al., 2026) shows that wrong labels can yield correct rewards for responses matching neither label nor reference. Agreement dominated by correct zeros need not supply a relative task signal. AEE measures joint target error and activity; reference replay checks changed or lost distinctions.

Verifier-level quantities such as Youden's index $J = \mathrm{TPR} - \mathrm{FPR}$ describe discrimination between correct and incorrect answers under a specified distribution (Rad et al., 2026). Our group-level, policy-dependent question is: *when does an incorrect target coincide with a potentially nonzero relative task signal?* Let B indicate an incorrect target and $H = \{\mathrm{Var}(r) > 0\}$ an active task-reward group. Define

$$E = \mathbb{P}(B \cap H) = \mathbb{P}(B)\mathbb{P}(H \mid B), \qquad \mathrm{ACR} = \mathbb{P}(H^c). \tag{1}$$

This is *Active Error Exposure* (AEE). ACR measures exact collapse; threshold-based diagnostics (He et al., 2026) also count small reward variance. At equal label-error rates, AEE can differ with policy, target representation, or group size.

Measure the joint event. Error and activity need not be independent:

$$E = \mathbb{P}(B)\mathbb{P}(H) + \mathrm{Cov}(\mathbf{1}_B, \mathbf{1}_H). \tag{2}$$

Estimation requires correctness annotations aligned with sampled groups. Appendix I provides a minimal group-level estimator. Section 4.5 describes our reference construction and paired replay.

3.2 Reachability and the task-update contribution

Target reachability and reward activity. Let $r_1, \ldots, r_m$, $m \geq 2$, be i.i.d. binary task rewards conditional on a fixed question and target, with ρ the policy's positive-reward probability. Then

$$\mathbb{P}(\mathrm{Var}(r) > 0 \mid q, \tilde{y}) = 1 - \rho^m - (1-\rho)^m. \tag{3}$$

This identity also appears in effective-group rollout allocation (Voss et al., 2026). For correct or incorrect targets, activity peaks at intermediate ρ and vanishes at either extreme, including a dominant wrong target, connecting wrong-target reachability (Shao et al., 2025) to group-relative feedback. Here ρ is current own-target acceptance probability, not the frozen discovery score used for ADPO difficulty scaling.

Exposure separates frequency from direction. For group-centered task advantages with a positive denominator,

$$A_i(r) = \frac{r_i - \bar{r}}{s(r) + \epsilon}, \tag{4}$$

constant rewards give $A_i = 0$. At a fixed policy, consider the task-only on-policy score surrogate

$$U = \frac{1}{m}\sum_{i=1}^{m} A_i(r)\nabla_\theta \log \pi_\theta(a_i \mid q). \tag{5}$$

Advantages are detached; expectations use the current group-sampling law. Since $U = 0$ on H^c,

$$\mathbb{E}[\mathbf{1}_B U] = E\, v, \qquad v = \mathbb{E}[U \mid B, H]. \tag{6}$$

The right side is zero when $E = 0$. Exposure gives the frequency of potential wrong-target contributions; v carries conditional magnitude and direction. Activity does not guarantee a nonzero vector; equal AEE need not imply equal updates or outcomes. See Appendix I for the proof and distinction from optimizer updates.

Zero task advantage is not unchanged supervision. On the same trajectories, a wrong target can yield $(0, 0)$ instead of the correct target's $(1, 0)$, removing a useful distinction despite zero noisy advantage.

3.3 Two routes to changed feedback

Active noisy rewards can change reinforcement; collapsed rewards can erase reference distinctions, a route invisible to E. On fixed trajectories, paired replay compares reward vectors and centered task proxies, reporting reward transitions, strict proxy sign flips, and mean $|A_i(r_{\text{train}}) - A_i(r_{\text{ref}})|$. A mixed reference group becoming constant is a *lost reference distinction*. These task proxies omit full ADPO scaling, KL and clipping. Appendix I adds the reference contribution lost on collapsed noisy groups to Equation 6.

Exposure evolves with the policy. Both E_t and v_t change across steps. Task-only accumulation involves $\sum_t \eta_t E_t v_t$, not just $\sum_t E_t$: learning-rate weights, conditional magnitudes, and cancellation between directions matter. Equation 3 suggests a local test: hold the questions and policy fixed, change same-question target reachability, and inspect activity *after* substitution. Section 4.5 combines this test with semantic review.

4 Experiments

4.1 Experimental Setup

The models complete two curriculum/executor rounds, ending at step100; R2 runs reuse each method's R1 and R2 dataset. We report all three paired R2 seed runs, s1–s3, at a 4,096-token output cap, followed by a paired consistency summary. The s1 checkpoint-resumption/reference exception is documented in Appendix A. The saved-response diagnostics use s2, the pair used for the stopping and truncation audits. Among the three recorded R2 pairs, s2 is the clean diagnostic pair: it has no resume exception and uses the same inference stack within each benchmark. s1 includes the checkpoint/KL-reference exception, while s3 is cross-stack (L/P on MATH and P/L on Olympiad). We report all three runs, but use s2 for the saved-response and mechanism diagnostics rather than treating their spread as seed-only variability.

We use tool-free chain-of-thought evaluation with temperature 1.0, top-p 1.0, 256 samples per question, and a 4,096-token output cap. Evaluations at 2,048 tokens are reported separately in Appendix B. The datasets are MATH-500, a fixed 500-question subset of a local English Olympiad mathematics collection, and all 30 AIME25 questions. AIME25 was used in training-time validation, not an untouched holdout. Its nine checkpoint runs share one evaluation stack, with paired question seeds within each R2 repeat; evaluation RNG differs across repeats. Exact prompts and data provenance are in Appendices A and N. The primary extraction rule selects the last boxed answer; first-boxed extraction provides a sensitivity check on the same saved responses.

For c correct answers among n samples (256 here; 64 in the longer-output check in Section 4.4), we use the estimator (Chen et al., 2021)

$$\widehat{\text{pass@}}\,k(q) = 1 - \frac{\binom{n-c}{k}}{\binom{n}{k}}, \qquad 1 \le k \le n. \tag{7}$$

At $k = n$, this records whether at least one saved answer is correct, which we call *finite-budget coverage*. We report percentages and percentage-point differences.

All Agent0 entries denote our reproduction. Base/R1 and R2 are presented as separate evaluation groups. We inspect each paired difference and report its three-run mean and sample standard deviation to summarize consistency across runs. Identification intervals reported below are bounds, not confidence intervals.

Table 1: Main accuracy and coverage results: cap4096, 500 MATH/Olympiad questions, 256 samples per question, last-box grades. R2 rows also report first-box pass@256 gaps (pp) for Oracle-based minus Agent0. AIME25 is reported separately in Appendix N.

	MATH-500 last-box (%)			Olympiad last-box (%)			First-box gap (pp)	
Model	@1	@32	@256	@1	@32	@256	MATH	Olympiad
Base	36.39	–	96.80	16.97	–	77.40	–	–
Agent0 R1	58.12	–	96.00	27.35	–	75.40	–	–
Oracle-based R1	64.58	–	96.80	28.48	–	78.00	–	–
s1 Agent0	63.13	90.43	93.80	30.04	58.56	67.40	–	–
s1 Oracle-based	66.88	93.07	96.40	30.10	64.37	78.40	–	–
s1 first-box gap	–	–	–	–	–	–	+2.20	+9.40
s2 Agent0	63.56	90.30	94.20	29.83	58.13	67.80	–	–
s2 Oracle-based	67.70	93.03	96.60	30.87	64.06	77.00	–	–
s3 Agent0	64.41	90.30	93.40	30.09	57.86	66.00	–	–
s3 Oracle-based	66.86	93.07	97.60	29.91	63.80	77.00	–	–
s2 first-box gap	–	–	–	–	–	–	+2.00	+7.40
s3 first-box gap	–	–	–	–	–	–	+4.40	+9.80
Mean first-box gap	–	–	–	–	–	–	$+2.87 \pm 1.33$	$+8.87 \pm 1.29$

4.2 Accuracy and Coverage Across Rounds

Oracle-based R1 improves Base pass@1 from 36.39% to 64.58% on MATH-500 and 16.97% to 28.48% on Olympiad (Table 1). Its pass@256 matches Base on MATH (96.8%) and is 78.0% versus 77.4% on Olympiad: higher accuracy accompanies high aggregate coverage. Agent0 R1 also improves pass@1, with lower coverage in this comparison. Appendix A retains the early-checkpoint comparison, recovery records and bounds for its one unresolved grade.

Oracle-based R2 has higher pass@256 in all three pairs on both primary 500-question benchmarks (Table 1). The paired gaps are 2.60/2.40/4.20 points on MATH and 11.00/9.20/11.00 on Olympiad. Single-sample gains are less consistent: Olympiad s3 slightly favors Agent0 at pass@1 despite its positive coverage gap. In s2, the Olympiad gap grows from 1.03 at $k = 1$ to 3.21 at $k = 8$, 5.93 at $k = 32$, and 9.20 at $k = 256$. In the separate 2,048-token three-run evaluation, Olympiad pass@1 is nearly tied (29.96% versus 30.03%), while the coverage gap is 8.60 points (Appendix B). Figure 2 separates Base/R1 and R2 protocols across all 256 sampling budgets.

We also record AIME25 as a third benchmark in Appendix N: it contains 30 questions and was used for training-time validation, not as a clean holdout. The three R2 coverage gaps are positive, but Oracle-based covers 16/15/13 questions versus Base's 17/30 (R1 covers 16). Thus AIME25 supports relative retention in these runs, not expansion over Base; its small-sample standard deviations are descriptive.

4.3 Consistency Across Seeds and Extraction Rules

All three paired coverage differences remain positive under both extraction rules on MATH and Olympiad (Table 1). Mean last-box gaps are 3.07 ± 0.99 points on MATH and 10.40 ± 1.04 on Olympiad. AIME25 is a separate 30-question training-time validation benchmark, not a clean holdout; its full results and small-sample variability are reported in Appendix N. Accuracy is more sensitive: in s2 on MATH, last-box pass@1 is 63.56% versus 67.70%, while first-box gives 71.85% versus 67.32% for Agent0 and Oracle-based, respectively. On the same responses, changing extraction reverses the single-sample ranking while retaining the coverage gap (Jo et al., 2025). Thus the coverage result is more stable across these two extraction rules than the single-sample comparison. Section 4.4 examines output length and stopping.

4.4 Output Budget and Coverage

Output length and stopping. Agent0 reaches the 4,096-token cap on 85.92% of MATH and 87.27% of Olympiad responses, versus 0.37% and 1.10% for Oracle-based. AIME25 shows the same contrast: 88.54% versus 1.22% (Table 2). Oracle-based outputs are shorter on average under the same sample and output caps.

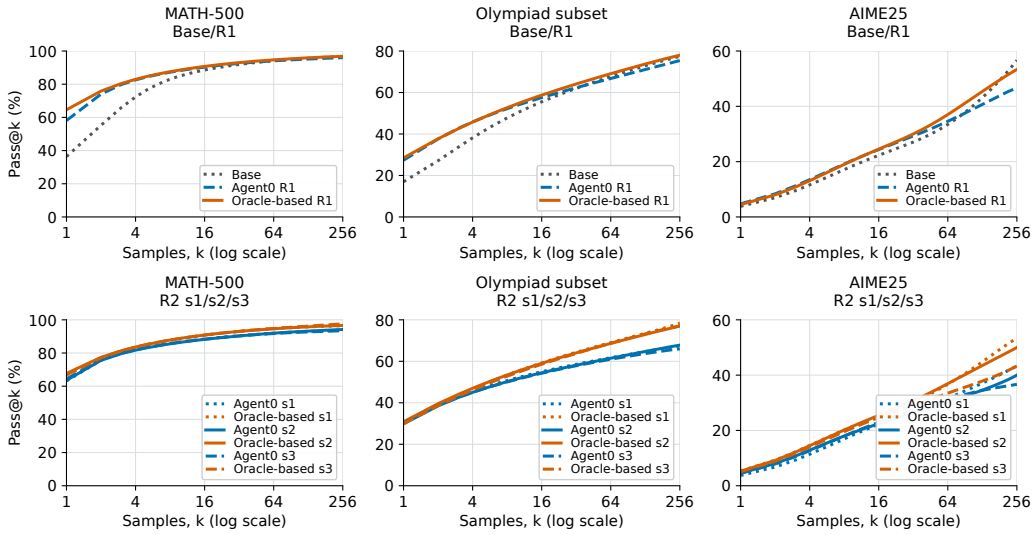

Figure 2: Coverage across budgets: Base/R1 (top) and all R2 seed runs (bottom, individual s1–s3 curves). Each curve uses 500 MATH/Olympiad questions or 30 AIME25 questions, with 256 responses per question. Lines show individual runs; y-axis ranges differ by column.

Table 2: Output behavior for the 4,096-token s2 diagnostic pair. Length denotes termination at the output cap, including responses with earlier answers. Means are per response.

	MATH-500		Olympiad subset		AIME25	
Method	Tokens	Length (%)	Tokens	Length (%)	Tokens	Length (%)
Agent0	3769.7	85.92	3825.1	87.27	3861.8	88.54
Oracle-based	539.7	0.37	789.7	1.10	966.8	1.22

Saved-token check. Neither candidate EOS is followed by continued generation in the 40,960 audited responses (Appendix G). This excludes ignored-EOS continuation within that sample; the analysis below examines length stops across the MATH/Olympiad s2 responses.

Saved-response truncation diagnostics. Across all 512,000 saved MATH/Olympiad s2 responses, no question has fewer than 50% Agent0 length stops; among 259 eligible Olympiad questions, conditional pass@32 is 58.27% versus 61.96% (a 3.69-point gap), but this selected analysis is neither a correction nor a lower bound for the full-set 9.20-point pass@256 gap because budgets, questions and retained response distributions differ (Appendix J); we therefore test a larger cap without filtering.

Coverage under a longer output budget. Native 8K and 16K runs use the same hash-frozen 200-question Olympiad subset, the same s2 checkpoints, 64 responses per question, and within-run 4K token-ID prefixes (Appendices K and L; Table 3).

Persistence does not establish equivalence or length independence: the 16K−4K change interval contains zero, and 31.14% of Agent0 responses remain capped.

Efficiency is part of coverage. The native 16K comparison makes the cost dimension explicit: the Oracle-based checkpoint obtains higher pass@32 while emitting roughly 12.6 times fewer tokens per response. More budget is not a monotone answer-quality guarantee for Agent0: from the same 4K prefix to the 16K response, 276 last-box decisions are rescued but 409 change from positive to negative. These transitions are grader decisions, so we use them to characterize answer instability under additional decoding; see Section 4.5 for the grading interpretation.

Table 3: Native-output budget summary: last-box pass@32 gaps are Oracle-based minus Agent0 (percentage points).

Budget	Last-box gap (pp)	Gap change vs 4K	Agent0 capped
4K prefix	7.24	–	–
8K	7.34	+0.10	53.09%
16K	7.28	−0.04, [−0.62, 0.55]	31.14%

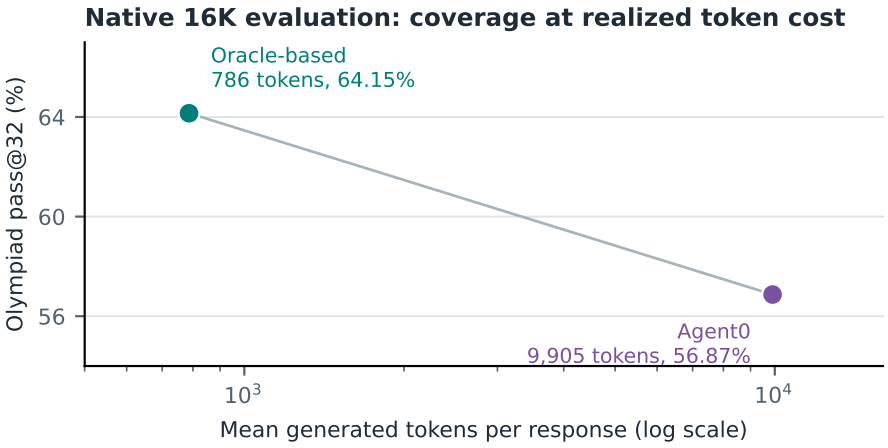

Figure 3: Native 16K Olympiad evaluation on the same fixed 200-question subset. The figure reports pass@32 against realized mean output tokens, not only the nominal cap. Oracle-based reaches 64.15% at 786 tokens per response, while Agent0 reaches 56.87% at 9,905 tokens. This is a descriptive efficiency–coverage contrast under a shared evaluation budget and is not a matched-training-cost causal estimate. Both checkpoints use the same evaluation budget (200 questions $\times$ 64 responses per question $\times$ a 16K cap); token differences reflect policy behavior rather than evaluation-configuration differences.

4.5 Auditing Delivered Feedback

4.5.1 Frozen model-assisted references

The review pilot contains 72 inferred groups of 16 saved trajectories, covering 63 distinct statements. Given limited expert-review capacity, we use `gpt-6-astra` for model-assisted mathematical review. Candidate answers and derivations are frozen before comparison with training labels and trajectories. Forty statements receive candidate answers; the remaining 23 include ambiguity, inconsistent or missing conditions, a placeholder, and a multipart task. The review record retains these cases with their dispositions. All transition counts below are grader decisions, not expert-verified correctness.

Replaying the training grader reproduces all 1,152 stored task rewards. Candidate-answer filtering leaves 12 Agent0 groups and 32 execution-grounded groups for raw reference substitution. Eligibility and blocked samples are documented in Appendix C.

4.5.2 Target errors and representation errors must be separated

Simply substituting reference-answer strings changes 67 rewards from 1 to 0 for Agent0 and 126 for the execution-grounded route. The latter number is driven in part by equivalent decimal and symbolic representations receiving different grader outputs.

We therefore add a post-hoc representation control. When the frozen reference and training target are judged mathematically equivalent, the intervention retains the original target encoding; other-

Table 4: Heldout same-question feedback probe on the fixed 12-task mask. Fixed-policy replay uses the same 768 responses and 48 groups per target arm. Hit rate is the fraction accepted against that arm's target.

Target	Hit rate (%)	All-negative	All-positive	Mixed
Source reference	20.44	8	0	40
Reachable candidate	25.39	7	0	41
Low-reach candidate	0.65	43	0	5

wise it substitutes the reference. Numerical tolerances and symbolic judgments are disclosed in Appendix C. The aggregate table there also excludes two interpretation-sensitive cases from the respective routes.

For Agent0, four of eleven eligible groups change their centered proxies, including 20 trajectories with strict sign flips. For the eligible execution-grounded groups, retaining equivalent targets produces no intervention change by construction. The following cases expose the distinct interfaces behind these aggregate changes.

Two further cases (an incorrect target that changes which trajectories are rewarded, and an incorrect target whose centered signal is unchanged) are given in Appendix C.

Correct target, defective reward distinction. A regular hexagon of side four has area $24\sqrt{3}$. All sixteen answers pass the review's equivalence checks, but the decimal target rewards only two. A radical target changes accepted forms without accepting every equivalent answer: correct targets need not yield correct reward distinctions.

4.5.3 A HELDOUT SAME-QUESTION FEEDBACK INTERVENTION

A separate probe fixes 40 DeepMath tasks and one Oracle-based R1 policy. Discovery sampling selects two candidate targets from each eligible question's own answers, then freezes a 12-task substitution mask before fresh validation sampling. Only target fields differ across the exported datasets. The same 768 heldout responses on masked tasks are scored under source, reachable, and low-reach targets and partitioned into 48 groups of 16 (Appendix E).

Reachable and low-reach targets yield 41 versus 5 mixed groups (Table 4), an active-fraction gap of 75.00 points. Their hit-rate gap is 24.74 points. Thus the intended *grader-reachability* contrast survives fresh sampling. Low-reach substitution also turns 35 of the 40 source-mixed groups all-negative, versus six under reachable substitution. Low observed activity can therefore coincide with loss of a reference distinction, exactly the second route omitted by AEE alone.

Semantic review tests the intervention itself. A subsequent blinded model-assisted review identifies one exactly equivalent low-reach target and four other interpretation concerns on the mask (Appendix H). In the volume case, the source fraction and the target string `\pi / 12` denote the same scalar but receive seven and zero hits out of 64. Apparent low reachability can thus arise from the representation–grader interface without mathematical corruption. The original 12-task results remain intact; a seven-task post-hoc sensitivity slice is reported separately. Together, the controlled probe and its review support checking both reward activity and reference semantics. The fixed-response comparison isolates how terminal targets change delivered task signals.

A separate exploratory target-only pair completed 40 steps per arm on 22 shared questions. Tool-rollout replay changes all 18 observed conflict-group reward vectors, but evaluation terminated with main estimates withheld (Appendix M).

A calibration-informed target-only pair completed 40 steps per arm on 248 shared questions. This intervention gives a clear training-side result: changing only the conflict target changes the delivered group-relative feedback. Reconstruction across all steps, including zero-learning-rate step 1, found 190/5,120 active group exposures under E and 115/5,120 under M. On conflict groups, E is nearly all-positive (2,435/2,542), whereas M is nearly all-negative (2,537/2,542); both endpoints suppress group-centered task advantages as predicted by Equation 3. E's 107 active groups are concentrated early (85 in steps 1–10), while M has only five across all 40 steps.

The downstream benchmark effect of this intervention is a separate question: the available grading versions leave it unidentified, so we report identification bounds rather than a point estimate. The reconstructed quantities are pre-mask sequence scalars and active-group exposures, not effective examples, token advantages or gradients (Appendix M.1).

Implications and scope. Semantic target repair and representation-sensitive grader repair are distinct. Constant rewards need not imply bad supervision; reducing AEE can miss lost reference distinctions. Establishing intervention effects requires downstream evaluation and replication. Expert-adjudicated references, additional model families and more rounds are needed beyond this two-round, single-family study, with accuracy, coverage and cost tracked jointly.

4.6 Limitations

The paper makes three deliberately bounded claims. First, the label-only pair identifies a change in delivered group-relative feedback, not a downstream ability effect. Second, the main target/grader comparison still has unresolved records: P2 is a completed sensitivity regrade with 2,017 main-policy unknowns, versus 1,047 under the original last-box policy; the higher count reflects stricter parser acceptance, not data damage. The dependency-only control was cancelled after 720 fresh grades and is excluded. A single unified terminal regrade with near-zero unknowns remains needed before replacing the identification bounds with a point estimate. Third, the co-evolution results use one model family and recorded R2 runs with protocol differences: s1 has a resume/reference exception, s3 is cross-stack, and s2 supplies the clean diagnostic pair. Conflict semantics are AI-assisted with zero human-confirmed rows. The formal target-only pair has one seed and is not a common-rollout experiment. These limits define the scope of the efficiency, coverage and feedback claims rather than negate the observed training-side contrasts.

5 Related Work

Self-evolution, executable supervision and noise. Agent0 and Self-Questioning Language Models study endogenous supervision (Xia et al., 2025; Chen et al., 2025); executable supervision spans self-play, checks, environment synthesis and learned evaluation (Zhao et al., 2026; Zhou et al., 2026; Gao et al., 2026; Shi et al., 2026; Liao et al., 2026). GRPO originates in DeepSeekMath (Shao et al., 2024); related noise studies examine reward inversion, incorrect targets and trajectory corruption (Lv et al., 2025; Zhu & Kang, 2026; Rad et al., 2026). Wrong-target reachability (Shao et al., 2025), label–reward disagreement (Zuo et al., 2026) and group activity (He et al., 2026; Voss et al., 2026) motivate our AEE and paired-replay diagnostics for lost reference distinctions.

Coverage, extraction and local supervision. Low-k accuracy and high-k coverage can diverge (Chen et al., 2026; Zhou, 2026), while extraction can change rankings (Jo et al., 2025); we evaluate finite-budget co-evolution across output budgets and extraction rules. Distillation distinguishes aggregate teacher quality from locally usable supervision on fixed student responses (Li et al., 2026; Zhang et al., 2026); our probe instead varies terminal targets to inspect group-relative feedback.

6 Conclusion

Oracle-based co-evolution improves single-sample accuracy over Base while retaining high MATH/Olympiad coverage. All three R2 pairs favor it over Agent0 in finite-budget coverage on MATH and Olympiad under both extraction rules; the AIME25 appendix reports a more limited retention result. The advantage persists in Olympiad native 8K/16K checks. AEE and paired replay distinguish active erroneous supervision, lost reference distinctions and representation-sensitive grading. The label-only intervention establishes that target selection changes delivered group-relative feedback, while its downstream effect remains unidentified under the available grading versions.

AI USE STATEMENT

During manuscript preparation, we used generative AI tools to assist with writing and language polishing, including improvements to wording, grammar, clarity and organization. We also used these tools to assist with the retrieval and discovery of related work, including developing search terms, identifying candidate papers and locating relevant research.

AI-generated text and retrieval suggestions served as supporting material, not as independent scientific evidence. The authors are responsible for verifying the sources, citation information and technical statements incorporated into the manuscript, and take full responsibility for the content, accuracy and research conclusions of the final paper.

ETHICS STATEMENT

This study concerns mathematical benchmarks and model-generated tasks. Erroneous references and automated graders can distort both training and evaluation; we retain unresolved cases and distinguish model-assisted judgments from human adjudication. The MATH-500 preparation code identifies `HuggingFaceH4/MATH-500`; the Olympiad subset is reconstructed from a local English mathematics collection. The local AIME25 parquet declares `yentinglin/aime_2025`; it was also used in training-time validation. These benchmark instances serve evaluation; their problem texts are excluded from the planned release. Appendix A documents provenance and unresolved upstream bindings. We did not perform a full benchmark-contamination audit and do not claim exclusion from all model training.

REPRODUCIBILITY STATEMENT

Appendix A describes prompts, subset reconstruction, training configuration, checkpoint resumption and evaluation versions. Appendices B and F retain all runs and original summaries. Hash-pinned manifests and deterministic builders bind the tables and curves to saved grades without regrading. Diagnostic selection and replay are specified in Appendices C–E; mathematical assumptions and derivations are given in Appendix I. Appendices J–L give the saved-response truncation and native 8K/16K prefix analyses, including source-binding and selection limitations. Both native analyses retain saved full last-box grades and derive prefix and first-box grades with the pinned grader. Appendix M records the historical exploratory pair and its withheld main estimates; Appendix M.1 separately reports the formal pair's all-step task-scalar reconstruction, unresolved semantic review, and original-policy identification bounds. Appendix N records all nine AIME25 evaluations and the three-seed supplement, with hash-pinned report-to-table and per-question curve checks. The appendix grouping is only a navigation aid; historical section labels are retained so that frozen manifests and existing cross-references remain stable. Upon acceptance, we will release code and de-identified evaluation manifests, excluding benchmark problem texts and other materials without redistribution rights.

REFERENCES

Lili Chen, Mihir Prabhudesai, Katerina Fragkiadaki, Hao Liu, and Deepak Pathak. Self-questioning language models. *arXiv preprint arXiv:2508.03682*, 2025.

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

Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? *Advances in Neural Information Processing Systems*, 38:57654–57689, 2026.

Jiaxuan Gao, Jiaao Chen, Chuyi He, Shusheng Xu, Di Jin, and Yi Wu. From self-evolving synthetic data to verifiable-reward RL: Post-training multi-turn interactive tool-using agents. *arXiv preprint arXiv:2601.22607*, 2026.

Xixiang He, Qiyao Sun, Ao Cheng, Xingming Li, Xuanyu Ji, Hailun Lu, Runke Huang, and Qingyong Hu. Advantage collapse in group relative policy optimization: Diagnosis and mitigation. *arXiv preprint arXiv:2605.21125*, 2026.

Hwiyeol Jo, Joosung Lee, Jaehone Lee, Sang-Woo Lee, Joonsuk Park, and Kang Min Yoo. Finding answers in thought matters: Revisiting evaluation on large language models with reasoning. *arXiv preprint arXiv:2510.14773*, 2025.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huanang Gao, Wenkai Yang, Zhiyuan Liu, et al. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe. *arXiv preprint arXiv:2604.13016*, 2026.

Wenjie Liao, Liangjie Zhao, and Zehong Cao. UnifiedPlayers: Enhance tool-integrated reasoning in agentic reinforcement learning. *arXiv preprint arXiv:2609.20089*, 2026.

Ang Lv, Ruobing Xie, Xingwu Sun, Zhanhui Kang, and Rui Yan. The climb carves wisdom deeper than the summit: On the noisy rewards in learning to reason. *arXiv preprint arXiv:2505.22653*, 2025.

Ali Rad, Khashayar Filom, Darioush Keivan, Peyman Mohajerin Esfahani, and Ehsan Kamalinejad. Rate or fate? RLV$^{\varepsilon}$R: Reinforcement learning with verifiable noisy rewards. *arXiv preprint arXiv:2601.04411*, 2026.

Rulin Shao, Shuyue Stella Li, Rui Xin, Scott Geng, Yiping Wang, Sewoong Oh, Simon Shaolei Du, Nathan Lambert, Sewon Min, Ranjay Krishna, et al. Spurious rewards: Rethinking training signals in RLVR. *arXiv preprint arXiv:2506.10947*, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Yucheng Shi, Zhenwen Liang, Kishan Panaganti, Dian Yu, Wenhao Yu, and Haitao Mi. Learning to build the environment: Self-evolving reasoning RL via verifiable environment synthesis. *arXiv preprint arXiv:2605.14392*, 2026.

Elena Voss, Marcus Hale, and Sofia Reyes. Early verdicts, better budgets: Sequential adaptive rollout allocation for compute-efficient RLVR. *arXiv preprint arXiv:2607.26253*, 2026.

Peng Xia, Kaide Zeng, Jiaqi Liu, Can Qin, Fang Wu, Yiyang Zhou, Caiming Xiong, and Huaxiu Yao. Agent0: Unleashing self-evolving agents from zero data via tool-integrated reasonin. *arXiv preprint arXiv:2511.16043*, 2025.

XiuYu Zhang, Wei Chow, Junfeng Fang, Zhenkai Liang, and Tat-Seng Chua. What does privileged information add to on-policy self-distillation? *arXiv preprint arXiv:2609.20612*, 2026.

Andrew Zhao, Yiran Wu, Tong Wu, Quentin Xu, Yang Yue, Matthieu Lin, Shenzhi Wang, Qingyun Wu, Zilong Zheng, and Gao Huang. Absolute Zero: Reinforced self-play reasoning with zero data. *Advances in Neural Information Processing Systems*, 38:105816–105879, 2026.

Todd Zhou. When RLVR shrinks the reasoning boundary: Diagnosing pass@k inversion. *arXiv preprint arXiv:2607.20543*, 2026.

Yifei Zhou, Sergey Levine, Jason Weston, Xian Li, and Sainbayar Sukhbaatar. Self-challenging language model agents. *Advances in Neural Information Processing Systems*, 38:113959–113991, 2026.

Yuxuan Zhu and Daniel Kang. Noisy data is destructive to reinforcement learning with verifiable rewards. *arXiv preprint arXiv:2603.16140*, 2026.

Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Xinwei Long, Ermo Hua, Biqing Qi, et al. TTRL: Test-time reinforcement learning. *Advances in Neural Information Processing Systems*, 38:131459–131483, 2026.

A Protocol and Reproducibility Details

Placement of executable checks. UnifiedPlayers (Liao et al., 2026) trains an Evaluation Player to construct statement-conditioned verifiers whose verdicts directly reward executor trajectories. L4′ instead screens a candidate target offline; it is not the executor's reward function. EvoEnv (Shi et al., 2026) synthesizes reusable environments and reviews their source, rendered instances, and references before admission. Here the unit is a statement/program pair, and the fresh implementation receives the statement rather than the generating program or target. AReaL-SEA (Gao et al., 2026) generates per-instance checks for interactive tool environments and evolves synthesis and evaluation plans. Our loop instead updates curriculum and executor policies.

Evaluation prompt. The tool-free CoT system prompt is:

> You are a helpful assistant good at solving math problems. Reason step by step, then put your final answer within `\boxed{}`.

In the three-run 2,048-token matrix, it is rendered with the checkpoint tokenizer's chat template and an added generation prompt. The R2 trace matrix in Appendix F and the Base/checkpoint evaluations instead use fixed ChatML rendering with the shared base tokenizer. The trace runner stores rendered prompts, responses, extracted answers, grading outcomes and failures, output lengths, finish reasons, and parameter metadata. Count audits validate each question index, sample count, and available answer mapping. They do not independently certify the mathematical grader.

Olympiad subset. The local English mathematics collection has 674 records. Reconstructing the builder's nonempty-answer and statement-length-below-3,000-character rules leaves 672; sampling 500 with Python `Random(0)` reproduces the saved parquet's ordered question/answer content. Record positions and content hashes are saved. This establishes correspondence to local data, not a verified upstream dataset revision or permission to redistribute it. We consistently refer to a subset rather than the complete benchmark.

Evaluation protocols. The count-audited checkpoint comparison in Appendix B has broader baseline coverage but lacks full per-run execution metadata. The three-run 2,048-token matrix includes saved traces and both methods' final repeats; the complete R2 trace matrix in Appendix F reports its output caps and dataset-specific stacks separately. The runner-source difference for Agent0 repeats 2/3 in the three-run 2,048-token matrix is removal of the optional `truncate_prompt_tokens=None` argument. Those runs also changed from vLLM 0.9.1 / PyTorch 2.7.0 to vLLM 0.20.0 / PyTorch 2.11.0 (with a hardware-specific build), so that small source difference does not establish equivalent generation across stacks. For the three-run 2,048-token matrix, the asset builder recomputes all 256 points from 500 question counts per checkpoint/dataset and checks them against the saved curves. Its complete input consists of 20 checkpoint/dataset combinations, 10,000 question rows, and 2,560,000 sampled answers, including early checkpoints. These are recomputed saved counts, not new inference.

Base/checkpoint evaluation. Table 5 includes all eight 500-question model evaluations in the 4,096-token checkpoint audit: 4,000 question–checkpoint records and 1,024,000 responses. These comprise 3,951 original records and 49 records in a separately sealed recovery namespace; 1,023,999 grades are known Booleans and one remains null. A read-only CPU aggregator validated the saved envelopes against pinned plans, campaign contracts, dataset bytes, completion maps, and prior shard-acceptance records. The full table and manifest retain all eight evaluations and their uncertainty bounds. The original failed shard and controller remain failed, and no recovery record is written into the original campaign. All 256 pass@k points and identification bounds are retained with source hashes and per-question counts. First-box accuracy is unavailable for this evaluation. Table 1 selects pass@1/pass@256 for Base and both R1 models on MATH/Olympiad from this frozen table, adding AIME25 from Appendix N; the separate three-seed evaluation is summarized in Table 1.

Recovery and the unidentified grade. A grader stall interrupted Olympiad shard3 for Agent0 R2 step20. Its first 14 records were preserved; the remaining 49 question batches were regenerated

Table 5: Complete 4,096-token checkpoint evaluation (%), cap4096, 500 questions and 256 responses per cell. Parentheses give steps. Base/Agent0 R1 ran locally; Oracle-based R1/Agent0 R2 step20 ran remotely. † includes 49 separately regenerated questions and one unresolved grade; outward-rounded intervals are identification bounds. Pass@256 is exact.

	MATH-500				Olympiad subset			
Model	pass@1	pass@8	pass@32	pass@256	pass@1	pass@8	pass@32	pass@256
Base	36.39	83.00	91.89	96.80	16.97	47.74	62.01	77.40
Agent0 R1 (100)	58.12	87.22	92.51	96.00	27.35	52.21	62.33	75.40
Oracle-based R1 (100)	64.58	87.57	93.04	96.80	28.48	52.70	64.12	78.00
Agent0 R2 s1 (20)†	60.34	86.72	91.81	95.40	[28.3554, 28.3563]	[51.5604, 51.5661]	[60.2713, 60.2862]	70.60

† applies only to the Olympiad evaluation in the marked row.

under a separate plan with unchanged model, prompt, per-position seed, n256, cap4096, and context6144. The new attempt redistributed questions across eight workers and persisted raw outputs before grading. An isolated, unmodified grader received a three-second deadline per distinct answer/target pair. This is a disclosed regeneration and grading-policy change, not numerical identity with the lost original batch. At position115, sample41's expression `1**5**5**5**5` times out against target `-5,7`; its grade and aggregate correct count stay null, alongside four known-correct answers. The recovery status remains `generation_complete_grading_incomplete`. † Separate 49-question regeneration; one unresolved grade at position 115, sample 41, with four known-correct samples at that position. Intervals are tight identification bounds, not confidence intervals. Correct samples: [36295, 36296]/128000; pass@1: [28.3554, 28.3563]%; pass@256 is exactly 353/500. For every k, the exact upper-minus-lower width on the probability scale is $\binom{251}{k-1}/(500\binom{256}{k})$ (zero for $k \geq 253$). Interval endpoints rounded outward to four decimals. Null was not graded false; the original shard and controller remain failed.

Base/checkpoint environment and comparison scope. Both hosts record Python 3.11.13, vLLM 0.10.2, PyTorch 2.8.0, Transformers 4.57.6, and tokenizers 0.22.2 (`d1-eval-bw`), but have distinct host kernels and campaign contracts. The L20Y R2 step100 pairs instead use vLLM 0.9.1, PyTorch 2.7.0, and Transformers 4.52.4. The checkpoint evaluation uses the shared mathruler 0.1.0 grader-source hash, temperature/top-p of 1, top-k of -1, 4,096-token output cap, and 6,144-token context limit. Declared stop lists are null with `ignore_eos=false`; the shared tokenizer declares EOS 151643. The seed schedule matches R2 repeat 1, not every conditional repeat. Base and Agent0 R1 use the local host; Oracle-based R1/Agent0 R2 step20 share the remote host, subject to the recovery change above. Remote launch explicitly waived the remaining natural-stop qualification and n256 capacity pretests; these checks are not represented as passed or supplied by an old receipt. These common settings do not establish cross-machine/version equivalence. The audit checks saved model-inventory and environment bindings, not current weight contents or installed grader packages; weight provenance remains stat-only. GPU identity is not established by this checkpoint audit.

Executor configuration and training budgets. Saved configurations agree on ADPO, actor learning rate 10^{-6}, 10 warmup steps, weight decay 0.01, train/minibatch size 128, microbatch size 1, 16 rollouts, temperature/top-p 1, four tool turns, prompt/response limits 1024/4096, observation/action limits 512/2048, clipping 0.2, KL coefficient 0.01 with `low_var_kl`, and 100 training steps. R1 starts from the same Base path; R2 normally starts from each method's own R1 step100 checkpoint. The R2 training sets contain 4,637 Agent0 and 4,536 Oracle-based records. Their questions, targets, and difficulty metadata are curriculum outputs; the common data conversion carries both target and score into ADPO training. Equal step and batch settings do not imply equal realized token counts or effective updates. Full generation/rejection costs, tool costs, and optimizer-consumed groups require further reconstruction from execution records.

Table 6: 2,048-token R2 step100 evaluation, mean ± sample SD over three recorded conditional runs, including s1. Mixed inference stacks; 500 questions and 256 samples per question. Each cell places the mean above ± SD.

	MATH-500				Olympiad subset			
Method	pass@1	pass@8	pass@32	pass@256	pass@1	pass@8	pass@32	pass@256
Agent0	64.05 ±0.58	85.79 ±0.09	90.53 ±0.05	94.20 ±0.35	30.03 ±0.09	50.19 ±0.09	58.07 ±0.16	67.73 ±0.76
Oracle-based	67.30 ±0.53	87.80 ±0.21	92.94 ±0.14	96.60 ±0.53	29.96 ±0.59	52.76 ±0.35	63.41 ±0.13	76.33 ±0.95

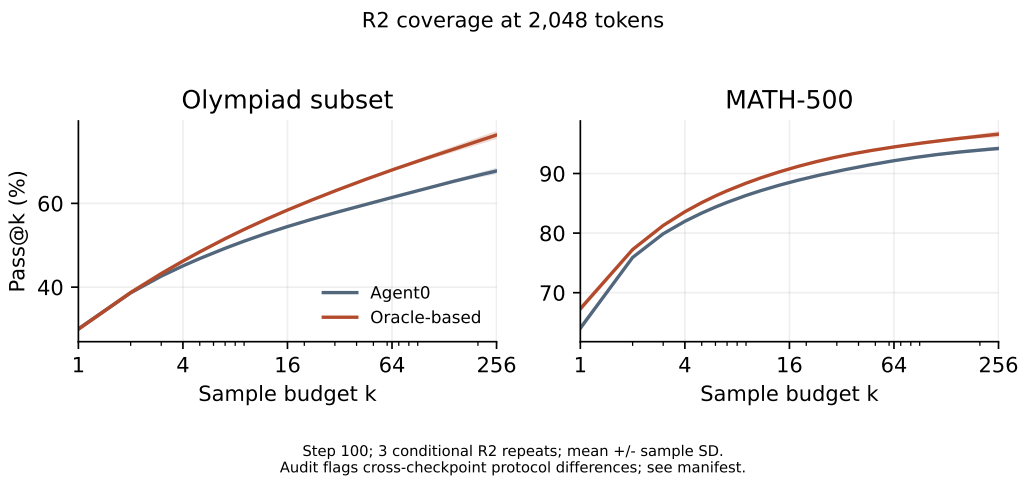

Figure 4: Three-run 2,048-token coverage curves. Lines are means and shading is sample SD; the protocol differences are described above.

Oracle-based R2 s1 checkpoint resumption. The initial configuration uses Oracle R1 step100 as its model path. The September1 resume configuration instead uses Base and resumes Oracle R2 step60; logs confirm loading the actor, optimizer, RNG, and learning-rate scheduler state. In the inspected implementation, a separate reference worker initializes from the configured model path, and checkpoint loading restores the actor but not that reference. Saved configuration and logs therefore indicate a Base KL reference for steps61–100 rather than the original Oracle R1 reference. This interpretation follows the available code path; the historical executable-source snapshot and reference weights were not independently frozen. Rollout GPU memory utilization also changes from 0.7 to 0.6. The saved s2/s3 configurations retain Oracle R1 and disable checkpoint resumption. All recorded runs remain in the results; the s1 difference is a training-protocol exception, not an identified cause of the coverage gains.

B Evaluation at Additional Budgets

Three-run 2,048-token evaluation. Table 6 and Figure 4 report all three conditional R2 runs, including s1. Sample standard deviations summarize the recorded training/evaluation histories, including the Oracle s1 resume exception, not seed-only variability. Agent0 s2/s3 use a different inference stack from the remaining 2,048-token runs. These descriptive summaries are not the 4,096-token comparisons in Table 1 and are not additional independent co-evolution experiments. The mean coverage gaps are 2.40 points on MATH and 8.60 on Olympiad. Olympiad pass@1 is 29.96% versus 30.03%; its coverage gaps grow from 2.56 points at $k = 8$ to 5.34 at $k = 32$ and 8.60 at $k = 256$.

Continuing from step20 to step100 reduces Agent0 s1 coverage by 1.8 points on MATH and 1.6 on Olympiad; Oracle-based mean changes are -0.27 and -0.93 points. Step100 is the fixed endpoint, not a test-selected checkpoint.

Table 7: Early-step comparison: within-run pass@256 changes at 2,048 tokens. Only Agent0 s1 has a step20 checkpoint in this evaluation matrix. These are checkpoint comparisons, not matched effective-update or token budgets.

		MATH-500			Olympiad subset		
Method	Repeats	Step 20	Step 100	Difference	Step 20	Step 100	Difference
Agent0	1	95.80	94.00	-1.80	70.20	68.60	-1.60
Oracle-based	3	96.87 ± 0.12	96.60 ± 0.53	-0.27 ± 0.46	77.27 ± 1.29	76.33 ± 0.95	-0.93 ± 1.21

Table 8: Output behavior at 2,048 tokens and step100, averaged over all three runs. Length denotes cap termination; multi-box counts syntactic occurrences.

	MATH-500			Olympiad subset		
Method	Tokens	Length (%)	Multi-box (%)	Tokens	Length (%)	Multi-box (%)
Agent0	1980.5	94.1	89.6	1995.9	94.9	90.1
Oracle-based	527.8	1.0	2.6	751.4	3.0	4.3

In s1, MATH first-box extraction changes Agent0 pass@1 from 63.61% to 71.86% and pass@256 from 94.0% to 94.4%; Oracle-based pass@1 changes from 67.02% to 66.63% with unchanged 96.8% coverage. Olympiad first-box pass@1 is 34.23% versus 29.50%, with coverage 68.2% versus 75.4%. These prefix scores reuse the saved responses rather than change generation.

Table 9: Count-audited checkpoint comparison (%), single-run values rather than repeat means. Per-run execution metadata are incomplete; these results are separate from Table 1. Standard RLVR uses externally labeled training data; it is a contextual baseline, not a zero-data method.

	MATH-500		Olympiad subset	
Method	pass@1	pass@256	pass@1	pass@256
Standard RLVR	54.53	96.60	25.98	76.20
Agent0 R1	59.24	96.40	27.67	75.00
Agent0 R2	63.38	93.60	29.94	67.00
Oracle-based R1	64.55	97.00	28.02	76.20
Oracle-based R2	66.94	97.20	29.69	76.60

Count-audited checkpoint comparison. The count-audited three-repeat R2 means for Agent0 versus execution-grounded training are 93.93 ± 0.31 versus 96.73 ± 0.42 on MATH-500 and 67.27 ± 1.03 versus 76.67 ± 0.31 on Olympiad at pass@256. The differences are 2.80 and 9.40 points. These are not additional independent training replicates of the trace-evaluation matrices.

Paired problems. For the R2 pair in the count-audited checkpoint comparison, execution-grounded training covers 20 MATH problems not covered by Agent0, while Agent0 covers two not covered by execution-grounded training; Olympiad counts are 55 versus seven. These are conditional observed any-success events, not estimates of permanent solution-support changes.

Early-checkpoint sampling repeats. A separate Agent0 step-20 evaluation reports MATH-500 pass@256 of 95.4, 96.0, and 95.6%. Their mean is 95.67%, with sample SD 0.31 points. Relative to its R1 reference at 96.4%, the difference is -0.73 ± 0.31 points. These sampling repeats are distinct from the within-training-repeat checkpoint comparisons in Table 7.

C Review and Reward-Proxy Details

The pilot requests eight groups per method at nominal steps 14, 74, 83, 90, and 94. The Agent0 step-74 file is blocked by ambiguous question extraction, leaving 32 versus 40 groups. Groups are inferred from compatible indices, prompts, labels, and trajectory identifiers; optimizer group

Table 10: Representation-controlled reward replay, conditional on model-assisted references. Transitions are grader-output changes, not human-verified false positive/negative rates. "Changed" counts groups with nonzero proxy changes. Oracle-based zeros follow from retaining equivalent target encodings; they do not certify the original reward vectors or excluded tasks.

Method	Groups / traces	$1 \to 0$	$0 \to 1$	Changed	Mean $\lvert\Delta A\rvert$
Agent0 R2	11 / 176	54	11	4	0.3465
Oracle-based R2	31 / 496	0	0	0	0.0000

UUIDs were not persisted. The model-assisted review covers 63 distinct statements: 40 candidate answers, three ambiguous, ten inconsistent, eight underspecified, one placeholder, and one multipart. Arithmetic checks support particular derivations, not population-level annotation accuracy.

The proxy uses $-1/+1$ task rewards, sample standard deviation, and $\epsilon = 10^{-6}$ in Equation 4. It omits ADPO scaling and the rest of the optimization computation. Reference answers are frozen before label/trajectory comparison, but representation-equivalence decisions were added after raw-format grader failures were observed. Numerical comparisons use relative tolerance 10^{-9} and absolute tolerance 10^{-12}; symbolic equivalence is model-reviewed. Neither is human gold certification.

The primary representation control excludes a logarithm-base ambiguity (task 17). The conservative Agent0 comparison also excludes task 60, whose fixed-k and maximum-over-k readings differ. Including its frozen fixed-k interpretation gives 12 groups/192 traces, 66 transitions $1 \to 0$, 14 transitions $0 \to 1$, five changed groups, 35 strict sign flips, and mean absolute proxy change 0.4456. Neither version yields a population error rate. No lost-clean-signal event was observed in this 63-statement comparison; the counterexample in Section 3 is analytical, not a frequency estimate. The separate pilot below measures loss of a *source-reference* reward distinction, without certifying that reference as mathematically correct.

Why not multiply marginal error and activity rates? In the main-text exposure definition (Equation 1), replacing the conditional probability with the overall active-group rate requires independence. That assumption is particularly inappropriate when the hypothesis is precisely that incorrect labels concentrate in collapsed groups. We do not use the product of marginal rates as a measured E or the joint event as a sufficient criterion for downstream harm.

D Curriculum and Target-Corruption Probes

Curriculum design probes. A 20-step probe retains program targets while replacing difficulty signal p with self-consistency s. Mean program-judged acceptance rates in the final datasets are 0.869 and 0.889, and 15.8% versus 12.9% lie in the $[0.3, 0.8]$ difficulty window. This single-run curriculum comparison has no corresponding downstream executor training or pass@k comparison.

In the L4′ ablation, both datasets contain 4,536 examples from a common generation pool. Removing the filter changes membership; 30.6% of the unfiltered set disagrees with the check. MATH pass@256 is 96.8% with the filter and 96.6% without, one run per setting. Filter disagreement is not an independently adjudicated label-error rate, and the small gap is not an equivalence result. A historical summary reports predominantly all-negative groups among 592 filter-disagreeing trajectories in 2,048 rollouts; the exact group numerator was not recovered. These observations motivate same-question feedback inspection rather than a label-only causal conclusion.

Six-run corruption pilot. A completed pilot trains six R2 runs with a nominal 30% target replacement rate, three repeats in each of two difficulty-selected conditions. The observed MATH-500 pass@256 values are 96.2, 95.8, 95.4% in one condition and 97.4, 97.0, 97.2% in the other: means 95.8 ± 0.4 and 97.2 ± 0.2%. We report absolute endpoints because the historical analysis notes do not consistently identify the reference for their deltas.

The intended experiment was to distinguish wrong targets that the policy can versus cannot readily sample. Its implementation draws replacement values from other questions' wrong-answer pools, checks string inequality rather than guaranteed mathematical nonequivalence, and changes task dif-

ficulty distributions. It therefore does not establish the intended same-question reachability contrast. The predicted clear degradation in the low-collapse condition was not established; the pilot cannot validate a causal exposure criterion. Filling the remaining cells of the same design would not repair this identification problem.

E SAME-QUESTION HELDOUT FEEDBACK PILOT

A distinct cohort and a fixed policy. This pilot uses 40 tasks from the local DeepMath training split, distinct from the 63-statement review and its 40 candidate answers in Appendix C. After schema checks and whitespace-normalized exact-question deduplication, tasks are hash-ranked with a fixed seed, before model sampling. Every arm retains all 40 tasks and the same execution-grounded R1 step-100 policy. The source-reference arm retains the dataset answers; these are not independently adjudicated gold. No arm is trained, and no downstream evaluation is performed for this intervention.

Discovery and target freezing. We generate 32 responses per task. Both candidate targets must occur among *that task's own* discovery answers and be rejected against the source reference, and against each other, by the training grader in both directions. The reachable candidate has observed hit frequency in $[4/32, 24/32]$; the low-reach candidate has frequency $1/32$. Fixed frequency and lexical rules break ties. Twenty-eight of 40 tasks have a qualifying pair. A fixed hash ranking selects 12 of them, freezing a common 30% substitution mask before validation. Only the target-answer field differs across the three exported datasets; row order, prompts, difficulty scores and other metadata are checked for identity after parquet round trips. The other 28 tasks remain unchanged. Candidate rejection is a grader outcome, not proof of mathematical error.

Independent sampling, shared replay. Validation draws 64 fresh responses for each of the same 40 tasks, using disjoint discovery/validation seed ranges. Labels are not supplied to the policy. Sampling is tool-free CoT at temperature 1.0, top-p 1.0, with a 2,048-token output cap. This yields 1,280 discovery and 2,560 validation responses, not three independently sampled validation sets. The same validation responses are replayed under each target with the training answer parser and correctness scorer, then partitioned in saved order into groups of 16. Table 4 includes every masked task: 12 tasks, 768 shared responses and 48 groups per arm. Hit rate measures acceptance against that arm's target, *not mathematical accuracy*.

What the manipulation establishes. The mean reachable-minus-low hit-rate gap is 24.74 percentage points; the mixed-group fraction differs by 75.00 points ($41/48$ versus $5/48$). These satisfy the predeclared descriptive checks: gaps of at least 0.10 in hit and active-group rates, a reachable active-group fraction of at least 0.50, and a nonzero source-to-reachable proxy change. They are engineering manipulation checks, not significance tests. Across all 40 tasks, source, reachable and low-reach targets yield 105, 106 and 70 mixed groups out of 160, respectively; the unmasked task outputs are retained.

Low-reach substitution changes 35 of the 40 source-mixed groups on masked tasks to all-negative rewards; reachable substitution changes six. Thus near-zero reward against a candidate can remove a distinction present under the source reference rather than leave its feedback unchanged. This is a source-relative diagnostic, not a certified loss of correct supervision or a measurement of training harm. The standardized task-reward proxies omit full ADPO, tool observations, KL, clipping and parameter updates.

Boundaries and audit. The single-policy, discovery-selected mask does not estimate a population error rate. The CoT distribution differs from the multi-turn tool rollout used in training; independent review of task semantics and candidate nonequivalence is still missing. Upstream scorer exception handling can swallow timeouts and return fallback negative grades, so rejection alone cannot establish nonequivalence. Passing this pilot does not establish training readiness, an effective-gradient mechanism, or safety of low-reach targets.

Discovery was interrupted by a command time limit after 34 tasks. The 1,088 completed responses were retained byte-for-byte; recovery generated only the remaining 192 responses with unchanged tasks and request seeds. Receipts bind the recovered trace, selection, validation and dataset hashes.

The paper's CPU builder checks these bindings, seed separation, label-only substitutions and arithmetic consistency of saved per-task reward summaries. It does not regrade responses or certify reference correctness.

F Complete R2 Trace Evaluation Matrix

Table 11 reports all fourteen completed R2 evaluation jobs: two Agent0 step-20 Olympiad evaluations at 2,048 tokens and twelve step-100 evaluations at 4,096 tokens. All jobs have passed full source and derived-record audits, covering 7,000 question–checkpoint evaluations and 1,792,000 generated responses. These are repeated evaluations of two 500-question subsets, not 7,000 distinct questions. This matrix is separate from the three-run 2,048-token evaluation in Appendix B.

Table 11: Complete R2 trace evaluation matrix (%). Each evaluated dataset cell group has 500 questions and 256 responses per question. Last is primary last-box grading; First is offline first-balanced-box prefix grading of the same responses. Oracle-based denotes the execution-grounded R2 method. All 14 completed jobs are included. L=L20Y/curriculum; P=PRO5000/curriculum-bw.

				MATH-500					Olympiad subset				
Method	Run	Step	Cap	Stack	Last @1	Last @256	First @1	First @256	Stack	Last @1	Last @256	First @1	First @256
Agent0	s2	20	2048	–	–	–	–	–	L	29.21	71.40	33.60	72.00
Agent0	s3	20	2048	–	–	–	–	–	L	29.18	71.20	33.83	71.60
Agent0	s1	100	4096	L	63.13	93.80	71.90	94.20	L	30.04	67.40	34.53	69.00
Oracle-based	s1	100	4096	L	66.88	96.40	66.52	96.40	L	30.10	78.40	29.97	78.40
Agent0	s2	100	4096	L	63.56	94.20	71.85	94.60	L	29.83	67.80	34.31	69.60
Oracle-based	s2	100	4096	L	67.70	96.60	67.32	96.60	L	30.87	77.00	30.72	77.00
Agent0	s3	100	4096	L	64.41	93.40	72.01	93.20	P	30.09	66.00	34.13	67.40
Oracle-based	s3	100	4096	P	66.86	97.60	66.43	97.60	L	29.91	77.00	29.79	77.20

– = not included in this evaluation matrix, not a zero or pending result. Stacks are dataset-specific; no cross-stack averaging.

Sampling uses fixed shared-base-tokenizer ChatML prompts, no tools, temperature 1, top-p 1, top-k -1, and one request of 256 samples per question. The 4,096-token cells use a 6,144-token model context, bf16, TP=PP=1, and no prefix caching. The saved L stack uses vLLM 0.9.1, PyTorch 2.7.0 and Transformers 4.52.4; P uses the PRO5000 vLLM 0.20.0 build, PyTorch 2.11.0+cu129 and Transformers 4.57.6. Both use mathruler 0.1.0, but the inference stacks are not assumed equivalent. The s3 pairs are cross-stack: Agent0/Oracle-based use L/P on MATH and P/L on Olympiad. Differences from the three-run 2,048-token matrix cannot be attributed solely to the token cap. The step-20 cells are separate from Table 7 and do not establish matched effective training budgets.

Source and reuse provenance. Existing evaluations were reused rather than rerun; the Agent0 s2/s3 step-20 MATH jobs were excluded from this matrix because results already existed. The cancelled duplicate campaign's manifests, logs, smoke tests and partial records remain preserved, but its partial results are not complete 500-question evaluations and are not included here. The source manifests and audit reports retain the evaluation identities and dataset-specific stack bindings; the 13-job audit archive is also preserved separately.

The CPU audit checks every saved source and derived record, question and seed identity, sampling parameters, rendered prompt, token counts, response and prefix hashes, and eight source plus eight derived completion proofs per job. It reuses saved grades without generating or regrading responses. The paper builder pins those audit reports, preserves all 256 points of both curves and the two contracts separately, and validates the table arithmetic. Hashes establish lineage, not mathematical truth; model inventories certify recorded metadata and shard statistics, not weight-file bytes. First-box analysis neither changes actual stopping nor measures saved tokens or latency. No new training, independent pipeline repetitions, or causal exposure test is represented by this table.

G Saved-Token Stopping Audit

The training launcher explicitly adds token ID 151645 to its EOS set, whereas the R2 trace evaluation requests no explicit additional stop IDs and uses the shared base tokenizer with EOS 151643. This motivated an audit rather than a silent change to the completed evaluation. We inspect positions 0–19 from each of the eight same-stack s1/s2 4,096-token jobs, retaining all 256 samples per question: 160 question-job records, 40,960 responses and 92,870,497 saved output IDs. This is a deterministic 4% subset of those jobs, not random sampling.

ID 151645 never occurs. ID 151643 occurs in 22,588 responses, always at the final saved position; neither target has a continuation after it. Of 18,379 length-finished responses, seven contain 151643 in the final cap slot, not before a continuation. Config/tokenizer hashes match the frozen inventory, and inspected record hashes, sample identities and token counts match completion proofs. The serialized sampling object is not an independent inspection of the engine's internal stopping state. No response is generated or regraded; no conclusion about unsampled trajectories, online savings, or the cause of the remaining nontermination follows from this check.

H Blinded Review of the Same-Question Pilot

This review concerns the 40-task pilot in Appendix E, not the 63-statement review in Appendix C. Two fresh-context model-assisted reviewers solve the 12 masked statements without source answers, candidate labels or rollouts; a third reviews the 28 unmasked statements. Their derivations are frozen before an unblinded comparison to the source and selected labels. All three reviewers use separate contexts of `gpt-6-astra`.

The comparison classifies 26 source references as likely supported, two as disagreeing under exact/general-claim readings, 11 as interpretation-sensitive, and one unresolved. These are review dispositions, not a population accuracy estimate. Among 24 selected candidate labels, 17 are judged nonequivalent under the recorded readings, one exactly equivalent, two equivalent under an endpoint convention, three interpretation-sensitive, and one unresolved. No candidate or reference is replaced in the frozen experiment.

An equivalent "low-reach" target. For the volume enclosed by $z = (x^2 + y^2)^2$ and $z = x$, cylindrical coordinates give

$$V = \int_{-\pi/2}^{\pi/2} \int_0^{(\cos\theta)^{1/3}} (r\cos\theta - r^4) r \, dr \, d\theta = \frac{1}{6} \int_{-\pi/2}^{\pi/2} \cos^2\theta \, d\theta = \frac{\pi}{12}.$$

The source LaTeX fraction and selected low-reach string `\pi / 12` (including its leading space) denote this same scalar. Nevertheless, their original saved heldout hit counts are seven and zero out of 64. This is a representation-sensitive target contrast, not semantic corruption. The source scalar and candidate scalar both omit the explanatory setup requested in the statement; that shared omission does not make the two scalar values different.

Interpretation and sensitivity. Four additional masked tasks concern an undefined endpoint of a right derivative, an unspecified matrix order, Boolean versus numerical zero, and a truncated candidate string. Excluding these four and the volume alias leaves seven tasks. On their *original saved grader summaries*, reachable/low-reach hits are 86/4 out of 448 and mixed groups 22/4 out of 28. The corresponding differences are 18.30 and 64.29 percentage points. This is a post-hoc descriptive slice, not a regraded experiment, a revised selection mask, or a newly passed training gate. The original 12-task table and all unresolved cases remain reported. Independent semantic adjudication, actual tool-rollout manipulation checks and common training initialization are still required before a training comparison.

I Exposure Decomposition and Its Scope

Estimating AEE and collapse from annotated groups. The following Python-style pseudocode estimates Equation 1 by equal-weight averaging over groups. Each reward vector belongs to one

question–target pair; `target_correct` annotates that target's mathematical correctness, not the correctness of an individual response. For a fixed group size $m \geq 2$, no binary-reward or within-group independence assumption is needed for this empirical count. Those assumptions are used only for the reachability formula in Equation 3.

```
def estimate_aee(rewards, target_correct):
    from math import isfinite
    G = len(rewards)
    assert G > 0 and len(target_correct) == G
    m = len(rewards[0])
    assert m >= 2
    exposed = collapsed = 0
    for r, correct in zip(rewards, target_correct):
        assert len(r) == m and type(correct) is bool
        assert all(isfinite(x) for x in r)
        active = any(x != r[0] for x in r[1:])
        exposed += (not correct) and active
        collapsed += not active
    return {"E": exposed / G, "ACR": collapsed / G}
```

The preconditions reject missing annotations, nonfinite rewards and unequal group sizes rather than silently dropping inputs. Activity means exact nonconstancy, not a small-variance threshold. For mixed group sizes, report size-stratified estimates; for selectively annotated groups, disclose annotation coverage and interpret $\widehat{E}$ as conditional on that selection. Repeated groups from one question each receive one group weight, so question-clustered uncertainty is appropriate when such groups are dependent. Report G, m, the sampling protocol and annotation provenance: the code does not validate the annotations. Neither estimate identifies update direction or training benefit, and a low $\widehat{E}$ does not rule out lost reference distinctions.

A group-level identity. Condition on the pre-sampling history $\mathcal{F}_t$, fixing the policy, question distribution, and target-generation law at step t. Let B_t indicate an incorrect target and H_t nonconstant noisy task rewards. For the integrable surrogate in Equation 5, use detached advantages with a positive normalizing denominator as in Equation 4. Then $U_t = 0$ on H_t^c, and

$$\begin{aligned}\mathbb{E}[\mathbf{1}_{B_t} U_t \mid \mathcal{F}_t] &= \mathbb{E}[\mathbf{1}_{B_t \cap H_t} U_t \mid \mathcal{F}_t] \\ &= \mathbb{P}(B_t \cap H_t \mid \mathcal{F}_t)\mathbb{E}[U_t \mid \mathcal{F}_t, B_t, H_t] = E_t v_t. \end{aligned} \tag{8}$$

Set the weighted conditional term to zero when its event has zero probability. No independence between error and activity is needed. This is a factorization of a stipulated task-surrogate contribution, not a claim that the surrogate is an unbiased gradient of expected reward or that we measured parameter gradients in the replay experiments.

The reference comparison has a second term. Recompute U^{ref} from reference rewards on the same policy and sampled group. Suppressing history conditioning, partitioning by noisy activity gives

$$\begin{aligned}\mathbb{E}[\mathbf{1}_B(U - U^{\mathrm{ref}})] = E\,\mathbb{E}[U - U^{\mathrm{ref}} \mid B, H] \\ - \mathbb{P}(B \cap H^c)\mathbb{E}[U^{\mathrm{ref}} \mid B, H^c]. \end{aligned} \tag{9}$$

The second term can be nonzero even when $E = 0$: a noisy collapsed group can suppress a reference task contribution. The identity is conditional on the stated reference, whose correctness must be separately justified. Reward and advantage replay measures parts of this comparison, not its parameter vectors.

Why a global verifier drift is not enough. In a separately assumed two-mode replicator model, let p be incorrect-mode mass, with $\dot{p} = p(1-p)(f_b - f_g)$. If f_b, f_g are expected group advantages, zero advantage outside H gives

$$f_b - f_g = \mathbb{P}(H \mid b)\mathbb{E}[A \mid b, H] - \mathbb{P}(H \mid g)\mathbb{E}[A \mid g, H].$$

Table 12: Coverage by fixed Agent0 length-stop bands. Empty low bands remain visible. Weighted contributions sum to the original 2.40 and 9.20 point gaps.

	MATH-500					Olympiad subset				
Stratum	Qs	Agent0 solved (%)	Oracle solved (%)	Gap (pp)	Contrib. (pp)	Qs	Agent0 solved (%)	Oracle solved (%)	Gap (pp)	Contrib. (pp)
Low	0	0 (NA)	0 (NA)	NA	0.00	0	0 (NA)	0 (NA)	NA	0.00
Mid	423	399 (94.33)	410 (96.93)	2.60	2.20	397	264 (66.50)	299 (75.31)	8.82	7.00
High	77	72 (93.51)	73 (94.81)	1.30	0.20	103	75 (72.82)	86 (83.50)	10.68	2.20

Original 256-answer coverage. Agent0 length-stop strata: low $[0, 0.5)$, mid $[0.5, 0.9)$, high $[0.9, 1]$. Contribution = question share times gap; empty strata remain NA with zero contribution.

The two activity probabilities need not equal each other or $\mathbb{P}(H)$. Consequently, multiplying an unconditional verifier-level fitness gap by $1 - \mathrm{ACR}$ is not a general finite-group correction. Moreover, a positive *good-minus-bad* gap decreases incorrect-mode mass, not the reverse. We use Equation 8 without asserting a new verifier-level convergence theorem or modifying the assumptions of Rad et al. (2026).

Across training steps. For predictable weights η_t, $\sum_t \eta_t E_t v_t$ sums the conditional mean wrong-target contributions of this surrogate along the training path. It is not generally the parameter displacement of an optimizer: KL and auxiliary losses, clipping, off-policy reuse, momentum, and adaptive scaling are outside the identity. Even at the surrogate level, equal unweighted $\sum_t E_t$ does not imply equal cumulative vectors, much less equal downstream coverage. These distinctions preserve the cross-round exposure question without promoting an unmeasured scalar threshold to a safety law.

J Saved-Response Truncation Analysis

Selection and audit. We analyze four s2 step100/cap4096 jobs: 500 questions per benchmark and method, 256 responses each, all on the L20Y stack. The 512,000 saved responses and last-box grades are read without generation, continuation or regrading; Agent0's fixed length-stop bands are $[0, 0.5)$, $[0.5, 0.9)$ and $[0.9, 1]$, including empty bands. These are analysis-time, not training-time, choices, and all benchmarks and bands are reported regardless of direction.

The builder validates question/target alignment between methods, record and response hashes, sample identities, token counts, finish reasons and known Boolean grades. Original per-question correct counts reconstruct all 256 pass@k points and saved aggregate length/token totals. There are 84 historically byte-pinned sources and 2,064 newly recorded byte snapshots, including the 2,000 raw records and 64 completion/shard records. The latter lack historical byte pins; their current bindings and aggregate agreement do not retroactively establish historical raw-byte integrity. Dataset bytes are pinned, but this analysis does not decode their rows again or retokenize answers.

What stratification shows. No question has an Agent0 length-stop fraction below 50%. Olympiad middle/high gaps are 8.82/10.68 points and MATH gaps 2.60/1.30, so differences are not confined to the at-least-90% band. The middle band remains substantially truncated, and conditioning on Agent0 output is not randomized; these are descriptive gaps, not significance claims or an exclusion of truncation.

Retained-response selection. Every original question remains in the per-question manifest with retained n and c. In this evaluation no question has zero retained responses; the analysis still represents zero-n cases explicitly as sample absence. Medians are 35 versus 256 retained responses on MATH and 32 versus 255 on Olympiad. Retained any-correct counts are 453 versus 483 and 290 versus 385 out of 500 questions. These events use unequal sample budgets and must not be read as natural-stop policy coverage or as an improved estimate of pass@256.

For each $k \in \{1, 8, 32\}$, we use only questions with at least k retained responses for *both* methods and apply Equation 7 with each method's own retained n, c. At $k = 32$, 328 MATH and 259

Table 13: Answers retained after excluding length stops. Per-question counts are unequal across methods; any-correct is not a matched-budget pass@256.

	MATH-500			Olympiad subset		
Method	Retained answers	Zero n	Any correct (%)	Retained answers	Zero n	Any correct (%)
Agent0	18023	0	453/500 (90.60)	16300	0	290/500 (58.00)
Oracle-based	127526	0	483/500 (96.60)	126590	0	385/500 (77.00)

	MATH-500	Olympiad subset
Method	n: min/Q1/median/Q3/max	n: min/Q1/median/Q3/max
Agent0	12/28/35/42/78	12/27/32/37/82
Oracle-based	229/255/256/256/256	187/254/255/256/256

All 500 questions per method. Retained = non-length stops. Any correct is NOT pass@256; zero n denotes sample absence. Full per-question n,c and distributions are in the manifest.

Table 14: Conditional pass@k on common eligible questions, using each method's retained n, c. Eligibility changes with k and is output-dependent.

	MATH-500				Olympiad subset			
k	Common Qs	Agent0 (%)	Oracle (%)	Gap (pp)	Common Qs	Agent0 (%)	Oracle (%)	Gap (pp)
1	500	63.33	67.81	4.49	500	29.56	31.00	1.45
8	500	85.58	87.88	2.30	500	50.13	53.53	3.39
32	328	92.31	94.62	2.31	259	58.27	61.96	3.69

Conditional pass@k uses each method's own retained n,c on the SAME eligible questions (both n at least k). Selected subsets are not an unbiased natural-stop distribution; do not compare across k subsets.

Olympiad questions qualify. Differences are 2.31 and 3.69 points. These conditional slices neither restore missing continuations nor estimate the distribution under an online stopping intervention. Different-k slices contain different questions.

Answers before the cap. We count responses containing the literal syntactic pattern `\boxed` followed by optional whitespace and an opening brace, separately from nonempty saved extractions and positive saved grades. An unfinished or repeated marker qualifies only as syntax. On Olympiad, 110,912 of Agent0's 111,700 length stops contain such a marker; 99,142 have a nonempty saved extraction and 33,365 a positive saved grade. The manifest retains all eight marker/extraction/grade joint cells. These observations reject the literal description that every capped response is answer-free, but cannot determine which capped failures would become correct with more tokens.

K Native 8K Same-Trajectory Coverage Check

Selection and generation. We retain the R2 s2 step100 checkpoints and select 200 of the existing 500 Olympiad questions before generating responses. Rows are sorted by the SHA256 of a canonical JSON object containing the dataset index, question hash and ground-truth hash; the first 200 are frozen for both methods. Selection does not use outcomes. Each method generates 64 responses per question with temperature 1.0, top-p 1.0, a maximum of 8,192 output tokens and context length 10,240. Both methods use the shared base tokenizer and fixed ChatML prompt, BF16, vLLM 0.10.2, and the same PRO5000 environment. Prefix caching is disabled. Seeds follow the recorded question-position rule, shared across methods. The 16 completed shards contain 400 question records and 25,600 responses.

Within-trajectory comparison. The shorter condition decodes exactly the first 4,096 saved output token IDs, not a character prefix. Naturally shorter responses remain shorter. There is no continuation of old 4K responses, independent 4K generation, or filtering by stopping behavior. Full

Table 15: Boxed markers, saved extractions and saved positive grades among length-stopped responses. Syntactic presence is not semantic correctness.

	MATH-500						Olympiad subset					
Method	Length stops	Marker	Stored box	True grade	Marker +true	All three	Length stops	Marker	Stored box	True grade	Marker +true	All three
Agent0	109977	109779	99639	69712	69712	69712	111700	110912	99142	33365	33365	33365
Oracle-based	474	53	49	5	5	5	1410	141	122	3	3	3

Counts of responses among length stops, not solutions. Marker = literal syntactic boxed marker; stored box = saved nonempty extraction; true = saved grade. All overlaps and marker occurrence counts are in the manifest.

last-box grades are retained from the runner; prefix last-box and prefix/full first-box grades are derived on CPU with the pinned mathruler package. Equation 7 uses $n = 64$ in every cell. Last-box pass@32 is the primary comparison; Table 16 retains all four sampling budgets and both extraction rules.

Table 16: Native 8K trajectories and their 4K token prefixes: pass@k (%), fixed 200-question Olympiad subset, 64 responses per question, R2 s2.

			Olympiad 200-question subset			
Extraction	Method	Output budget	@1	@8	@32	@64
Last	Agent0	4K prefix	31.09	51.16	57.13	59.50
Last	Agent0	8K full	30.20	50.95	56.66	59.00
Last	Oracle-based	4K prefix	31.73	53.58	63.55	68.50
Last	Oracle-based	8K full	31.76	53.61	63.55	68.50
First	Agent0	4K prefix	35.77	52.25	58.35	61.00
First	Agent0	8K full	35.81	52.30	58.37	61.00
First	Oracle-based	4K prefix	31.68	53.52	63.50	68.50
First	Oracle-based	8K full	31.70	53.55	63.51	68.50

Paired uncertainty. For each question, we compute the Oracle-based minus Agent0 pass@32 gap at each budget and the difference between those two gaps. We resample the 200 question clusters with replacement 10,000 times, sharing each draw across methods and budgets (seed 20260921). Table 17 reports the 2.5th and 97.5th percentiles with linear interpolation. These intervals describe question-level uncertainty for the fixed checkpoint pair, not variation across training seeds; secondary extraction and sampling-budget comparisons are sensitivity analyses, without multiplicity adjustment.

Table 17: Paired pass@32 gaps and gap changes (percentage points), with 95% question-bootstrap intervals. Gaps are Oracle-based minus Agent0.

	Olympiad 200-question subset			
Contrast	Last gap	Last 95% CI	First gap	First 95% CI
4K prefix gap	6.42	[3.59, 9.51]	5.15	[2.29, 8.28]
8K full gap	6.89	[4.01, 10.01]	5.14	[2.28, 8.27]
8K-4K gap change	0.47	[-0.48, 1.60]	-0.01	[-0.05, 0.00]

Additional tokens and answer revisions. Agent0 uses 2,781.75 more tokens per response beyond the 4K prefix on average, versus 12.01 for Oracle-based. Of 12,800 Agent0 responses, 277 move from a negative prefix last-box grade to a positive full grade, while 391 move in the reverse direction; Oracle-based has three and zero, respectively (Table 18). Full first-box pass@1 favors Agent0 (35.81% versus 31.70%), while pass@32 and pass@64 favor Oracle-based; the single-sample ranking remains extraction-sensitive.

Table 18: Native 8K output behavior. Length stops are percentages; last-box negative-to-positive (rescue) and positive-to-negative (harm) counts each use 12,800 responses per method.

	Olympiad 200-question subset			
Method	Mean tokens	Length stops (%)	Rescue	Harm
Agent0	6601.88	53.09	277	391
Oracle-based	785.90	0.04	3	0

Integrity and interpretation. All 16 original-runner completion checks pass. The analysis verifies the frozen selection, canonical record hashes, token counts, response hashes, saved grade/error fields, tokenizer files and grader source pins; reported grading errors are not converted into negative grades. Of 25,600 responses, 25,591 match batch token decoding exactly. Nine Agent0 responses differ only in trailing incomplete Unicode: token-by-token streaming decoding reproduces the saved text exactly, and both extracted boxes are unchanged. These cases are indexed in the report, without rewriting responses. Independent CPU checks reproduce point estimates, primary paired intervals and saved full-last counts; they do not constitute expert mathematical adjudication.

The observed coverage advantage is not removed by extending these trajectories from 4K to 8K. The interval for the gap change includes zero; length independence remains unresolved. With 53.09% of Agent0 responses still length-capped, this run alone leaves larger-budget effects open; Appendix L reports a separate native 16K run. This 200-question, 64-response estimand neither decomposes nor corrects the historical 500-question, 256-response gap, and does not identify an AEE-mediated training effect.

L NATIVE 16K SAME-TRAJECTORY COVERAGE CHECK

Protocol and paired prefixes. We retain the native-8K experiment's fixed 200 questions, s2 step100 checkpoint pair, tokenizer, ChatML prompt, sampling parameters and question-position seed rule, but generate 64 fresh responses per question with a 16,384-token output cap and 18,432-token context. The two methods run sequentially in the same PRO5000 environment. These are not continuations of the saved 8K responses. Shared seeds do not imply identical prefixes across native runs. The 4K and 8K conditions decode exactly the first 4,096 and 8,192 output token IDs of each new 16K trajectory; naturally shorter responses remain shorter. Every condition retains all 200 questions and 64 responses, without filtering length stops.

Table 19: Native 16K trajectories and their paired token prefixes: pass@k (%), fixed 200-question Olympiad subset, $n = 64$, R2 s2. These responses are separate from Table 16.

			Olympiad 200-question subset			
Extraction	Method	Output budget	@1	@8	@32	@64
Last	Agent0	4K prefix	30.82	51.10	56.91	59.00
Last	Agent0	8K prefix	30.02	50.89	56.82	59.50
Last	Agent0	16K full	29.78	50.84	56.87	59.50
Last	Oracle-based	4K prefix	31.82	53.73	64.15	69.50
Last	Oracle-based	8K prefix	31.84	53.75	64.15	69.50
Last	Oracle-based	16K full	31.84	53.75	64.15	69.50
First	Agent0	4K prefix	35.62	52.15	58.07	60.50
First	Agent0	8K prefix	35.63	52.15	58.07	60.50
First	Agent0	16K full	35.63	52.15	58.07	60.50
First	Oracle-based	4K prefix	31.75	53.65	64.11	69.50
First	Oracle-based	8K prefix	31.77	53.67	64.11	69.50
First	Oracle-based	16K full	31.77	53.67	64.11	69.50

Coverage and uncertainty. Full last-box pass@32 is 56.87% for Agent0 and 64.15% for Oracle-based, a 7.28-point gap; pass@64 is 59.50% versus 69.50%. We use Equation 7 and 10,000 paired question-bootstrap resamples (seed 20260921), sharing each draw across methods, budgets and extraction rules. Percentile intervals use linear interpolation. Table 20 reports all three budget gaps and gap changes; last-box pass@32 is primary. Secondary comparisons are sensitivity analyses without multiplicity adjustment. Intervals describe question uncertainty for one checkpoint pair, not training-seed variation. First-box pass@1 still favors Agent0 (35.63% versus 31.77%), while its pass@32 favors Oracle-based (58.07% versus 64.11%).

Table 20: Native 16K paired pass@32 gaps and gap changes (pp), with 95% question-bootstrap intervals. Gaps are Oracle-based minus Agent0; displayed zeros may reflect rounding rather than exact equality.

	Olympiad 200-question subset			
Contrast	Last gap	Last 95% CI	First gap	First 95% CI
4K prefix gap	7.24	[4.25, 10.47]	6.04	[3.07, 9.29]
8K prefix gap	7.34	[4.39, 10.52]	6.04	[3.08, 9.29]
16K full gap	7.28	[4.33, 10.48]	6.04	[3.08, 9.29]
8K-4K gap change	0.10	[-0.57, 0.63]	0.00	[0.00, 0.00]
16K-8K gap change	-0.06	[-0.22, 0.08]	0.00	[0.00, 0.00]
16K-4K gap change	0.04	[-0.62, 0.55]	0.00	[0.00, 0.00]

Tokens and revisions. Agent0 reaches the 16K cap on 3,986 of 12,800 responses (31.14%); Oracle-based has no length stops. Their mean lengths are 9,904.60 and 786.47 tokens. From the 4K prefix to the full response, Agent0 has 276 negative-to-positive last-box grade transitions and 409 positive-to-negative transitions; Oracle-based has two and zero. These are grader transitions, not adjudicated mathematical repairs or errors. Prefix truncation at budget B means length strictly

greater than B; a naturally stopped response of length exactly B is not counted as truncated. Full length stops require the saved length finish reason and exactly 16,384 tokens.

Table 21: Native 16K output behavior. Rescue/harm are last-box negative-to-positive and positive-to-negative transitions from 4K prefix to 16K full, out of 12,800 responses per method.

	Olympiad 200-question subset			
Method	Mean tokens	Length stops (%)	Rescue	Harm
Agent0	9904.60	31.14	276	409
Oracle-based	786.47	0.00	2	0

Integrity and scope. All 16 original-runner shard checks were rerun successfully. The analysis validates frozen question identities, canonical record and response hashes, sampling settings, token counts, completion receipts and tokenizer/grader source pins. Saved full last-box grades are retained; both prefix last-box grades and all three first-box grades are derived on CPU. Reported grading errors are not converted to incorrect answers. Of 25,600 responses, 25,589 match batch decoding; 11 trailing incomplete-Unicode discrepancies pass exact streaming reproduction with unchanged boxes. All 468 tracked input files pass drift checks. This does not certify checkpoint weight bytes, native dependency binaries, historical symbolic-library versions or expert mathematical correctness.

The last-box 16K-minus-4K gap change is $+0.04$ points with interval $[-0.62, 0.55]$; the 16K-minus-8K change is -0.06 with $[-0.22, 0.08]$. Thus extending these trajectories does not remove the observed advantage. Intervals containing zero are not equivalence tests, and 31.14% of Agent0 responses remain capped. This fixed200/$n64$ estimand does not correct the historical 500-question/$n256$ gap, establish matched realized token costs, or identify an AEE-mediated training effect.

M Exploratory Target-Only Training Ablation

Same questions, different targets. A separate single-seed exploratory pair starts both actor and KL reference from the same Oracle-based R2 step100 checkpoint, with fresh optimizers and no resume. The two arms use the same ordered 22 training questions: 11 supported execution/plurality conflicts and 11 audited agreements. On conflict questions, E retains the original execution-answer bytes, supported by statement-derived review; M uses the frozen discovery plurality representative. Only the target field changes between arms. Agreement targets, prompts, row order and discovery difficulty metadata are identical. The original heldout split retains six conflicts and six agreements. This is not the 40-task fixed-policy probe of Appendix E, nor the main co-evolution comparison.

Selection and qualification. The full model-assisted, statement-derived review of 95 selected conflicts classified 17 as supported conflicts, 35 as equivalent representations, 34 as ambiguous and nine as execution-incorrect. These counts concern the selected conflicts, not population label-error rates; the review is not independent expert certification. The retained training/heldout conflicts are 11/6. Protocol amendments changed the pool to R3/R3s/R4 and initialization to R2, and reduced discovery support and extraction thresholds. Discovery used CoT, 16 samples and a 2,048-token cap, whereas training uses tool rollouts. Frozen difficulty is historical execution-grader acceptance, not certified success probability. Conflicts and agreements are not difficulty-matched (training acceptance ranges 0–0.3125 and 0.9375–1); this limits conflict-versus-agreement comparisons, although E and M still train on identical questions. Raw-program repeatability was not newly established. The completion receipt therefore retains `protocol_qualified=false`, separately from successful training and verified paired controls.

Completed training and operative intervention. Each arm completed 40 steps with batch/minibatch size 16, 16 rollouts per prompt, learning rate 10^{-6}, ten warmup steps and KL coefficient 0.01. Training allows four tool turns with prompt/response limits 1024/4096 and observation/action limits 512/2048. It uses ADPO advantage estimation with a vanilla dual-clipped actor loss (clip 0.2, dual-clip constant 3), not adaptive actor clipping. Shared seeded shuffling with drop-last samples 16 of 22 questions each epoch, yielding 640 prompts and 10,240 initial rollout samples per arm. Shared seeds do not guarantee identical asynchronous execution or token-balancing order. Both arms retain checkpoints at steps 20 and 40; E's cleanup issue was resolved in separate receipts, not by retraining or resuming its optimizer.

Separately, shared E-policy tool-rollout smoke responses from steps 1–2 cover all 11 conflict questions in 18 step/question groups of 16. Replaying both targets changes all 18 conflict reward vectors (Table 22). Across all smoke responses, grader alignment is 512/512, with no recorded extraction, timeout or grading-error artifact. These are not separate E/M samples or step40 outcomes. Predominantly all-positive E and all-negative M groups show why a target-induced reward change need not supply a relative task signal.

Table 22: Exploratory target-only tool-rollout smoke check. Both targets replay the same E-policy responses; counts concern conflict groups, not downstream performance or separate trained-arm samples.

Target	Conflict groups	All-negative	All-positive	Mixed
E	18	0	15	3
M	18	17	0	1

Evaluation terminated; main estimates withheld. Generation is complete for Init/E40/M40 on MATH-500, the 500-question Olympiad subset and 12 heldout questions: 48,576 responses in total, with 16 responses per question and a 4,096-token cap. The declared primary comparison is Olympiad last-box `math_verify` pass@16, E40 minus M40; first-box and mathruler results are sensitivities, not interchangeable primary graders. The heldout comparison uses E targets for every model. This historical campaign terminated with `unknown_grades_main_withheld`; cleanup passed, while its success flag remains false. It is no longer grading. Unknown grades remain unknown rather than failures. The completed training and manipulation check establish neither down-

stream benefit nor harm or training-seed robustness. The formal pair below is a separate experiment, not a replacement of these records.

M.1 Formal target-only pair: activity and scoring limits

M.1.1 Design and paired control

Separate design and semantic scope. The formal pair uses 248 shared training questions (124 conflicts and 124 agreements) and 32 heldout questions (16 of each), with shared initialization and one paired training seed. Only conflict targets differ: E uses execution answers and M the frozen discovery numeric-plurality representatives. Each arm completed 40 steps with batch/minibatch size 128 and 16 rollouts per question. This calibration-informed amendment retains `protocol_qualified=false`; topic/length matching is not difficulty matching, and programmatic checks are not human semantic approval. It is separate from the 22/12-question pair above and from the main co-evolution comparisons.

M.1.2 Validity of the conflict subset

Independent arithmetic checks support all 124 training conflicts: 98 modular-inverse and 26 Fibonacci-modulo questions. This composition is a limitation of the selected conflict set, not a claim about the population of model errors. In 99/124 cases, the winning numeric cluster has at most half the samples. This is expected under the largest-cluster supervision rule: M deliberately represents Agent0-style numerical plurality, not an absolute-majority filter. The historical semantic-v3 screen flags 43 rows non-exhaustively. Assistant full-context review of all 968 winning responses labels 106 questions as unanimous asserted numeric winners, 10 as mixed-cluster context and 8 as unresolved under a conservative all-winner criterion. These are assistant judgments, not independent semantic truth: all 124 questions still require human review, with zero human-confirmed rows and no exclusions, relabeling or cleaned-vote recount of the original training data. The 43 screening flags are not equivalent to the 18 mixed or unresolved questions. Arithmetic correctness and numeric cluster membership alone do not establish asserted-answer meaning; assistant coverage completion is not human certification.

M.1.3 Training-side reward saturation

All-step task-scalar reconstruction. For terminal task reward $R_i = 2\,\mathrm{hit}_i - 1$, the audit reconstructs $a_i = w(R_i - \bar{R})/(s_R + 10^{-6})$, where s_R is the sample standard deviation and the positive ADPO scale w uses frozen discovery difficulty. These are pre-mask sequence scalars, not measured advantage tensors, token counts or gradients. Table 23 counts repeated step/question group exposures, not unique examples, across all 40 steps of each arm, including step 1.

Table 23: Formal pair, all 40 steps. Groups have 16 sequences. Active means mixed own-target rewards. Nonzero scalars are reconstructed pre-mask sequence exposures; the totals include zero-learning-rate step 1.

Arm	Stratum	Groups	All-negative	All-positive	Active	Nonzero scalars
E	Agreement	2578	0	2495	83	1328
E	Conflict	2542	0	2435	107	1712
E	All	5120	0	4930	190	3040
M	Agreement	2578	0	2468	110	1760
M	Conflict	2542	2537	0	5	80
M	All	5120	2537	2468	115	1840

Of 2,542 conflict-group exposures per arm, E has 2,435 all-positive groups and M has 2,537 all-negative groups; activity is sparse at both ends of the two-sided saturation curve. The resulting conflict active fractions are 4.21% and 0.20%, respectively. E's activity is front-loaded, while M is absent in 36/40 steps rather than absent at every step. The five positive M conflict samples occur in five mixed groups and are retained. Overall, E and M have 3,040 and 1,840 nonzero sequence-scalar exposures out of 81,920 per arm. These counts are not effective-example or gradient-volume estimates.

Equation 3 predicts low activity as $\rho \to 0$ and as $\rho \to 1$. The observed M all-negative and E all-positive conflict regimes therefore instantiate both endpoints in one paired intervention. The early E concentration is consistent with a rapidly exhausted feedback window when the execution target

Table 24: Conflict active-group exposures by training interval. Counts are repeated group exposures, not unique questions or independent samples.

Steps	E active	M active
1–10	85	3
11–20	13	0
21–30	5	0
31–40	4	2
Total	107	5

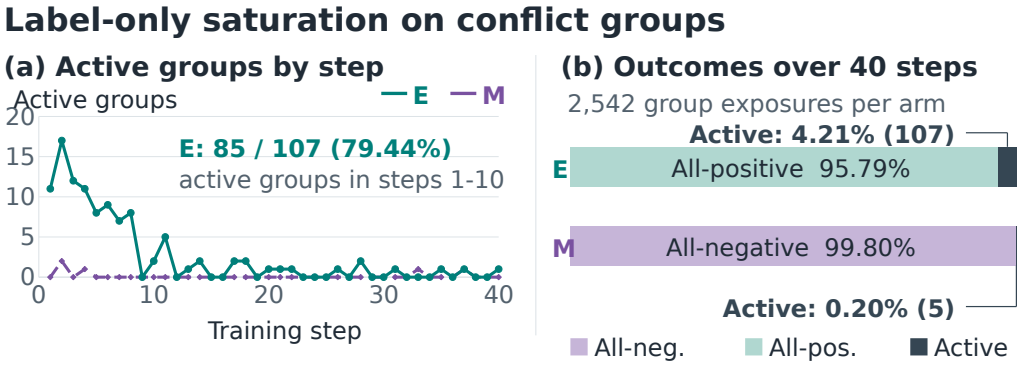

Figure 5: Conflict-group activity across the 40-step formal pair. Left: exact per-step active-group counts for the execution target (E) and numerical-plurality target (M). Right: the exact all-negative, all-positive and active proportions over 2,542 conflict-group exposures per arm. Active means mixed binary rewards within a group; exposures are repeated training records, not distinct questions, effective examples or gradients. E's early concentration and the two endpoint saturation regimes are shown without smoothing.

is highly reachable; it does not establish a larger gradient or a downstream gain. Reconstructed sequence-level scalars agree with recorded min/max values to within 1.29×10^{-7}; this is not a measurement of token-level advantages or parameter updates. Both arms have zero learning rate at step 1: all 2,048 paired prompts are identical, but only 1,771 responses are identical. These are paired requests, not a common-rollout experiment; step-1 gradients may still populate optimizer moments.

Only steps 10/20/30/40 have completion-bound raw records; the other 36 per arm are newly snapshot-hashed surviving records with consistency checks, not equal historical integrity. The receipts are local hash attestations. Exact policy-mask token counts and gradient decompositions are unavailable. Zero task advantage removes the direct reward-policy gradient, not KL contributions, shared-parameter effects, Adam momentum or weight decay. The observed saturation is consistent with Equation 3, not proof of gradient benefit or downstream mediation.

M.1.4 Downstream identification limits

Original-policy bounds; no replacement scores. Evaluation covers Init/E40/M40 on MATH-500, Olympiad-500 and the 32-question formal heldout, with 16 responses per question: 49,536 responses in total. Under the original last-box/`math_verify` policy, Olympiad pass@16 E40−M40 has identification bounds $[-12.4, +13.8]$ percentage points. These arise from unresolved scoring; the effect remains unidentified, without establishing superiority, inferiority, equivalence or absence of an effect. All 1,047 main-policy unknown samples remain unknown across datasets and models; 500 belong to the primary comparison (246 E and 254 M). Separate endpoint bootstraps do not form a joint 95% effect interval. The parser repair has been independently reviewed; versioned full regrading has not yet produced replacement results. No additional training is part of this analysis.

Table 25: Primary original-policy identification bounds for E40−M40 (percentage points). Every reported interval includes zero.

Version	Benchmark	pass@1	pass@8	pass@16
Original	MATH-500	[-1.01,+1.28]	[-2.31,+3.14]	[-3.0,+2.8]
Original	Olympiad-500	[-2.74,+3.51]	[-9.76,+10.05]	[-12.4,+13.8]

Table 26: P2 sensitivity only, not a completed scoring repair: identification bounds for E40−M40 (percentage points). These bounds are secondary and every reported interval includes zero.

Version	Benchmark	pass@16
P2	MATH-500	[-4.8,+4.6]
P2	Olympiad-500	[-16.0,+16.8]

The heldout 16+16 split is a secondary diagnostic: its conflict and agreement pass@1 endpoint intervals have bootstrap ranges crossing zero. The primary contrast therefore remains underdetermined under both grading versions. The available formal bounds do not support a claim of superiority, inferiority, equivalence, or absence of an effect.

M.1.5 Grading-version sensitivity

The original policy is primary because it was fixed before the formal result was observed; P2 is sensitivity analysis. P2 changes dependency loading and parser acceptance simultaneously, so it is not a completed "scoring repair." Its main-policy unknown count is 1,047 for the original and 2,017 for P2. The transition ledger is 164 original parse-unknown records becoming known, 1,118 original known records becoming parse-unknown, and 16 becoming timeout, for a net increase of 970 unknown records. The original 47 empty-box and five unbalanced-box records are retained, and 67 timeout records observed during the 16/32-worker campaign are disclosed rather than deleted or reclassified. A dependency-only control was attempted but not completed under a uniform protocol and is excluded.

M.1.6 Limitations

This is a single-seed, non-common-rollout intervention. The reconstructed quantity is a pre-mask sequence scalar rather than a token advantage tensor or gradient. Only steps 10/20/30/40 have completion-bound historical raw-hash receipts; the remaining step records are later snapshot-hashed records with consistency checks. Conflict semantics were AI-assisted and human-confirmed for zero rows. These limitations prevent seed-stability, common-rollout and human-certification claims.

N AIME25 Evaluation and Three-Run Replication

Scope and provenance. We use all 30 unchanged rows of the local `aime25_test.parquet`, whose metadata declares `yentinglin/aime_2025`, upstream split `train`. The dataset byte SHA256 is recorded in the pinned manifests accompanying the paper. An immutable upstream revision, redistribution rights and exclusion from model pretraining are not certified. This dataset was used in training-time validation; we do not call it an untouched holdout. A divergent local JSONL mirror was not mixed into this evaluation. Neither AIME24 nor AIME25 native 8K/16K runs are included.

Fixed evaluation contract. All nine checkpoint evaluations use the shared Qwen3-4B-Base tokenizer and the fixed ChatML prompt in Appendix A, without tools, checkpoint-specific templates or an injected thinking prefix. Temperature and top-p are 1; top-k is disabled. Each question receives 256 responses with a 4,096-token output cap and context limit 6,144. Native EOS remains enabled; generation does not stop at the first box. The common stack is vLLM 0.10.2, PyTorch 2.8.0, Transformers 4.57.6 and mathruler 0.1.0, with bfloat16 weights and prefix caching disabled. The nine evaluations contain 69,120 formal responses. The initial five-model campaign and four-checkpoint supplement contributed 38,400 and 30,720 respectively; their 40 and 32 smoke responses are excluded. The initial smoke plan's one-question description was stale: the frozen runner actually selected two questions per model.

Pairing and uncertainty. Base/R1 and both R2 s1 checkpoints use seed $20510922 + p$ at zero-based position p. The s2 and s3 pairs use $21510922 + p$ and $22510922 + p$. Thus methods share evaluation RNG seeds within a pair, but RNG realizations differ across repeats. R2 repeats reuse each method's R1 and R2 dataset; they are not independent full co-evolution pipelines or an isolation of training-seed variability. We retain all three pairs regardless of scores. The reported mean and sample SD use the three unrounded paired differences; no responses are pooled across seeds. One covered question is 3.33 percentage points. Three same-question repetitions do not create 90 independent test questions, and 256 responses do not create 256 independent problems.

Table 27: Complete AIME25 results (%): all nine evaluations, cap4096, n256. First-box extraction uses the same saved responses, not early-stopped generation. Covered columns report question counts, not necessarily nested solved sets.

	Last-box			First-box			Covered	
Model	@1	@32	@256	@1	@32	@256	Last	First
Base	3.84	27.07	56.67	3.84	28.01	56.67	17/30	17/30
Agent0 R1	4.74	29.24	46.67	5.49	29.97	46.67	14/30	14/30
Oracle-based R1	4.47	30.01	53.33	4.38	30.01	53.33	16/30	16/30
Agent0 R2 s1	3.72	26.86	43.33	4.11	27.57	46.67	13/30	14/30
Oracle-based R2 s1	5.35	30.28	53.33	5.31	30.03	53.33	16/30	16/30
Agent0 R2 s2	4.56	26.30	40.00	5.00	26.64	40.00	12/30	12/30
Oracle-based R2 s2	5.26	30.72	50.00	5.20	30.03	50.00	15/30	15/30
Agent0 R2 s3	4.47	27.39	36.67	4.88	28.60	36.67	11/30	11/30
Oracle-based R2 s3	5.08	29.14	43.33	5.07	28.82	43.33	13/30	13/30

Interpretation. Last-box R2 gaps at pass@1, pass@32 and pass@256 average 0.98 ± 0.56, 3.20 ± 1.35 and 8.89 ± 1.92 points; first-box gaps average 0.53 ± 0.58, 2.02 ± 1.63 and 7.78 ± 1.92. Every paired difference at these three budgets is positive under both extractors. This is a relative finite-budget advantage: Base covers 17 questions, compared with 16 for Oracle-based R1 and 16/15/13 for Oracle-based R2. Oracle-based R1 also trails Agent0 R1 at pass@1 under both extractors. The initial s1 last-box pass@256 question-bootstrap interval is $[-3.33, 23.33]$ points (10,000 paired question resamples); it is conditional on that run, not a three-run interval. We do not claim statistical significance from the sign or SD.

Output behavior. Across the nine AIME25 evaluations, mean output tokens and native length-stop rates were: Base 1,356.5/8.14%; Agent0 R1 3,243.6/62.77%; Oracle-based R1 975.1/1.08%; Agent0 R2 s1 3,971.1/93.37%; Oracle-based R2 s1 971.6/1.07%; Agent0 R2 s2 3,861.8/88.54%; Oracle-based R2 s2 966.8/1.22%; Agent0 R2 s3 3,993.8/94.08%; and Oracle-based R2 s3 989.8/1.18%. Native length-stop includes outputs containing earlier boxed answers; no low-truncation subset is selected.

Reproducibility and limits. The initial five-model report and the three-run supplement are retained separately; the supplement does not regenerate or overwrite s1. All 72 formal shards and both extraction records were checked, and the three-run summary was independently reproduced. The paper builder pins report bytes, recomputes all 256 pass@k points from per-question correct counts, and checks reported scores and output diagnostics without inference or regrading. Model metadata are hashed, whereas weight shards are checked by path, size and modification time, not weight-content hashes. Large length-stop differences remain. First-box extraction does not recover missing continuations, and the Olympiad 8K/16K checks cannot establish length independence on AIME25. These benchmark comparisons do not identify an AEE-mediated causal effect.