ORACLE COVERAGE CAN REVERSE RANKINGS: FINITE-BUDGET IDENTIFICATION IN TEST-TIME SCALING

Anonymous authors
Paper under double-blind review

ABSTRACT

Ranking systems by pass@K (oracle coverage) can disagree with ranking them by selected accuracy. We prove that strict oracle dominance at every sampling budget can coexist with plurality regret approaching one. We then derive sharp finite-budget identification bounds on plurality accuracy given correct-answer structure, abstention mass, and the largest wrong-answer probability. With these quantities fixed, extremal allocations of wrong-answer mass determine the accuracy range. Reanalysis of public response banks on 3,600 SuperGPQA problems finds high-budget ranking reversals between reasoning-effort settings after simultaneous correction, including when invalid responses participate in voting. On a 1,800-question development split at $K = 64$, adding the largest wrong-answer probability narrows empirical identification widths from 30.81–44.77 to 0.49–1.39 percentage points across three effort settings. These ranges describe information loss, not statistical confidence. We also give a fixed-score stopping certificate that skips verifier queries without changing the full weighted-vote decision. On 64 problems per policy, batch-preserving execution reduces verifier-stage time by 51–70%, excluding candidate generation. Code and artifacts are available at `https://anonymous.4open.science/r/oracle-selector-gap`.

1 INTRODUCTION

Additional test-time computation can improve reasoning performance (Brown et al., 2024; Snell et al., 2025; Muennighoff et al., 2025). Oracle coverage denotes pass@K: the probability that at least one of K sampled candidates is correct under the specified criterion. It is the accuracy of an oracle selector given candidate correctness labels (Chen et al., 2021). Practical systems instead use plurality voting, weighted self-consistency, or a learned verifier (Wang et al., 2023; Cobbe et al., 2021; Lightman et al., 2024). When a shared wrong answer has the largest probability mass, more samples can reinforce that answer instead of recovering oracle success.

What additional information makes selected accuracy identifiable at a finite budget? RLVR can improve single-sample accuracy (Shao et al., 2024; DeepSeek-AI, 2025); whether it expands reasoning boundaries or reweights existing trajectories remains debated (Yue et al., 2025; Liu et al., 2025a). We measure unrecovered coverage as

$$\mathrm{gap@}K = \mathrm{pass@}K - A_{\mathrm{maj}}@K, \tag{1}$$

where A_{maj} is plurality-vote accuracy over normalized final answers. The mapper defines the voting categories, so each gap depends on mapper, selector, and budget. Do oracle curves identify the best system for a named selector? Proposition 1 answers no even under strict dominance, without assuming plurality optimal or ruling out stronger selectors.

Our central result characterizes the exact range of plurality accuracy over all wrong-answer distributions consistent with retained categorical information. After fixing correct-category probabilities, abstention mass, and the largest wrong-answer probability, concentrating or dispersing the remaining wrong mass gives sharp finite-budget endpoints (Proposition 2). These endpoints define an admissible set of systems rather than forcing a ranking that the available information cannot justify (Figure 3).

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.

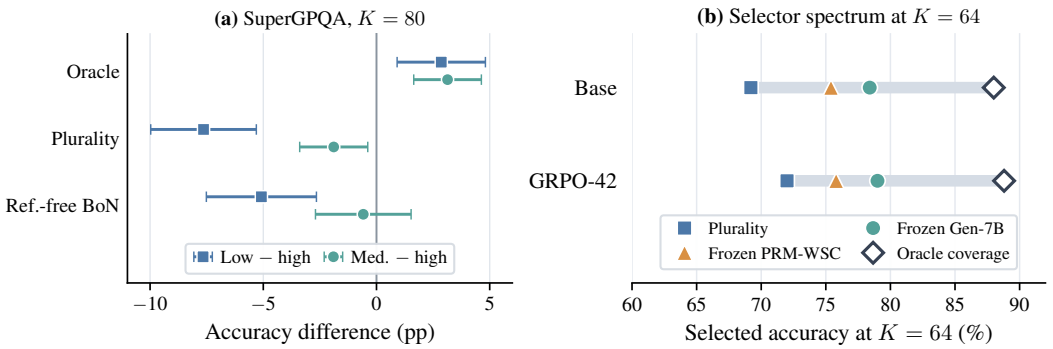

Figure 1: Selection depends on the evaluation contract. (a) On SuperGPQA, differences relative to gpt-oss-20b's high-effort setting show higher oracle coverage but lower plurality accuracy. Bars are simultaneous 95% intervals across all 144 contrasts; both crossings survive invalid-vote controls. The low–high reversal also holds for reference-free Best-of-N (BoN). These settings use different compute. (b) On MATH-500, practical selectors recover different portions of the same response bank's oracle coverage.

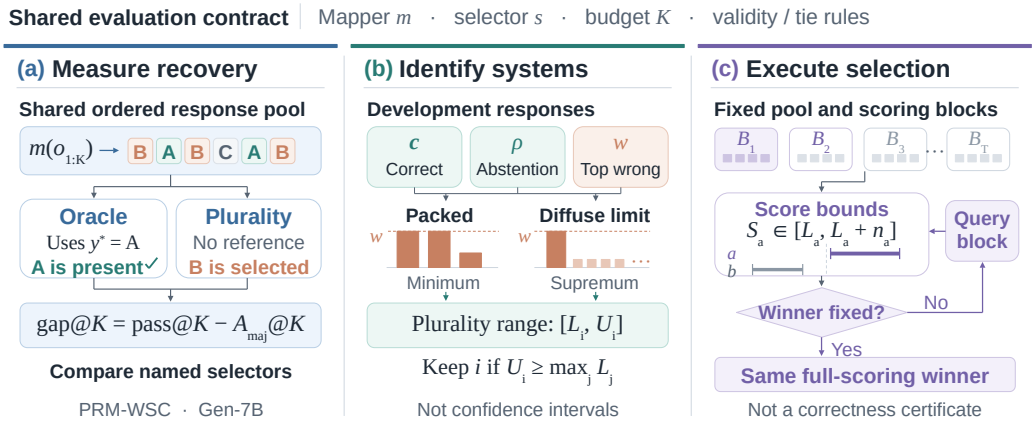

Figure 2: Selector-aware framework. (a) A correct answer can be present but lose the vote. (b) Correct-category probabilities c, abstention mass ρ, and largest wrong mass w define sharp identification ranges, not confidence intervals. (c) Fixed-block score bounds certify the full-scoring decision, not answer correctness.

Public response banks exhibit ranking reversals between reasoning-effort settings, though not a change in the overall winner when a stronger backbone is included. Controlled RLVR experiments establish a persistent residual, not a stable training-induced penalty. Disjoint-pool prediction tests the categorical account, and a fixed-score stopping certificate applies the same decision-preserving principle to verifier queries.

2 Framework

Selectors, mappers, and tie policy. Given i.i.d. responses $o_{1:K} \sim \pi_\theta(\cdot \mid x)$, the oracle succeeds if any extracted answer equals y^*. This upper bound applies to selection from the same candidate pool, not to generating new answers. Plurality returns the most frequent normalized answer, resolving finite-sample ties by first occurrence to match the released evaluator. The answer mapper and decision rule together constitute the operational selector: complete-box voting rejects incomplete boxes; structured-semantic voting additionally accepts explicit answer statements and exact rational equivalence; the released fallback is retained as a compatibility endpoint. At $K{=}1$, oracle, plurality, PRM Best-of-1, and weighted self-consistency are exactly identical; our tests enforce this invariant.

Three different questions. The residual in Eq. 1 differs from the remaining-error-normalized gain

$$G_K = \frac{A_{\text{maj}}@K - A_{\text{maj}}@1}{1 - A_{\text{maj}}@1}, \tag{2}$$

which accounts for the smaller headroom at higher pass@1. Neither quantity alone identifies a training effect. We compare GRPO with rejection-sampled SFT (RFT) at development-matched pass@1, but their data, objectives, and optimization paths differ. This is a decision reference, not a causal control; test oracle coverage never selects its checkpoint.

Oracle-induced deployment regret. For candidate systems under a fixed tie rule, let $i_{\text{orc}}(K)$ maximize pass@K and let $A_s^i(K)$ be the accuracy of deployed selector s. The decision cost of oracle-based model choice is $\mathcal{R}_s(K) = \max_i A_s^i(K) - A_s^{i_{\text{orc}}(K)}(K)$. It vanishes only when the oracle winner is also best for the named selector.

Categorical predictive account. For problem j, let $p_{*,j}$ be correct-answer probability, $p_{w,j}$ the largest eligible wrong-answer probability, and $\Delta_j = p_{*,j} - p_{w,j}$. The account concerns where categorical mass moves, not just per-problem correctness. We report normalized entropy, effective support $\exp(H)$, unique answers, top-wrong mass, and modal margin. Entropy alone cannot distinguish movement toward a correct answer from concentration on a shared wrong mode.

Gap as a missed-selection rate. For any fixed ordered pool, plurality correctness implies oracle success. Thus the problem-level gap is binary and admits the exact event representation

$$\text{gap}_x\,@K = \mathbb{1}[N_*(K) > 0,\ \mathcal{S}_{\text{maj}}(a_{1:K}) \neq y^*], \tag{3}$$

where $a_{1:K}$ are ordered mapped answers and $N(K)$ their counts. Same-pool measurement is descriptive; we instead estimate the histogram and test its predictions on disjoint responses. This aggregate account is not an instance-level selector or routing method; problem-level error is reported separately.

Given a population histogram p, oracle success has closed form $1 - (1 - p_*)^K$. Plurality success is

$$\sum_{n:\,\sum_a n_a = K} \text{Mult}(n; K, p)\tau_*(n). \tag{4}$$

Here $M(n)$ contains the eligible categories with maximal positive count, and $\tau_*(n) = \mathbb{1}[y^* \in M(n)]/|M(n)|$, or zero if $M(n)$ is empty. For i.i.d. draws, this is the exact conditional first-occurrence win probability. With several correct categories C, use $p_* = \sum_{a\in C} p_a$ and $\tau_C(n) = |M(n) \cap C|/|M(n)|$, again zero for empty $M(n)$. Correct categories retain separate votes; they are not merged using test labels. We average it over fresh counts from Pool A's histogram and compare the predicted gap with Eq. 3 on Pool B. No calibration coefficient is fitted.

2.1 From oracle impossibility to finite-budget identification

Proposition 1 (Oracle-curve impossibility and maximal regret)**.** *For any $q \in (0, 1/2)$ and $0 \leq \epsilon < 1/2 - q$, there are finite answer distributions P, Q with correct masses $q + \epsilon$ and q. At $\epsilon = 0$, their complete oracle curves are identical, yet any deterministic oracle-curve-only choice rule has worst-case limsup plurality regret one under relabeling; for randomized rules the minimax limsup expected regret is at least $1/2$. At $\epsilon > 0$, P has strictly higher pass@K for every K, so oracle selection always chooses P. Nevertheless, plurality accuracy under P converges to zero and under Q to one; hence $\mathcal{R}_{\text{maj}}(K) \to 1$.*

Put $q + \epsilon$ correct mass in P and concentrate its remainder on one wrong answer; in Q, spread the wrong mass over $m > (1 - q)/q$ answers. The strong law gives the opposite plurality limits; swapping their labels proves the minimax claim because the oracle input is unchanged. Proposition 5 also yields an explicit finite-K regret bound from the two modal margins. The appendix proves both claims and the countable-space extension. Hence the missing object is how incorrect mass is allocated, not merely whether correct mass is sampled. The categorical estimator in Eq. 4 supplies that information, and the disjoint-pool experiment tests whether it predicts the shortfall out of sample.

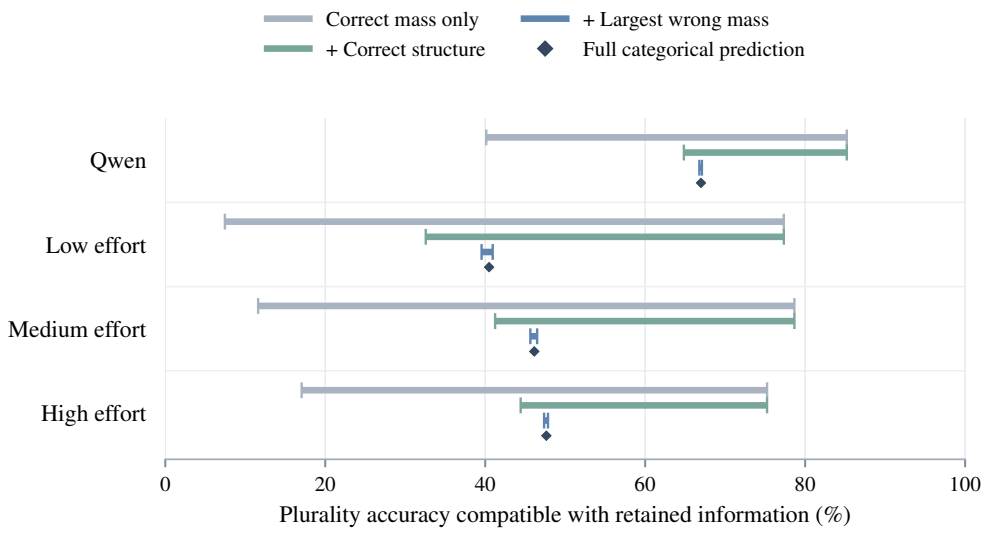

Figure 3: Information narrows the admissible decision set. At $K = 64$ on the 1,800-question SuperGPQA development split, retaining the largest wrong mass separates high from the other effort settings and Qwen from all three. Bars are empirical-distribution identification ranges, not confidence intervals; diamonds show full categorical predictions. Invalid outputs vote. Exact endpoints and estimation limitations appear in Appendix X.

Proposition 2 (Sharp finite-budget identification of plurality accuracy)**.** *Fix a finite vector of correct-category probabilities c, abstention mass ρ, and total eligible wrong mass $W = 1 - \sum c - \rho > 0$. For i.i.d. sampling and first-occurrence plurality ties, suppose the largest eligible wrong mass is $w > 0$. Among all such distributions, the minimum accuracy at each budget K packs W into $\lfloor W/w \rfloor$ atoms of mass w and one remainder. The supremum retains one atom w and splits $W - w$ into arbitrarily many equal atoms. With at most $m > 1$ wrong categories, the maximum instead uses $(w, (W - w)/(m - 1), \ldots, (W - w)/(m - 1))$.*

Balancing two wrong masses cannot hurt correct plurality: conditional on other counts, its win probability increases toward a balanced split. This specializes classical majorization reasoning (Rinott, 1973); Appendix X supplies the proof, exact finite-polynomial calculation, and correctness-only endpoints. It preserves distinct correct spellings and separates abstention from competing wrong answers.

Average the per-problem endpoints to obtain $[L_i, U_i]$ for each system. Discard i only if $U_i < \max_j L_j$; a strict winner satisfies $L_i > \max_{j \neq i} U_j$. On the 1,800-question SuperGPQA development split, correct-structure information alone leaves all three effort settings admissible at $K = 64$, with range widths 44.77/37.42/30.81 points. Adding the largest wrong mass reduces these to 1.39/0.85/0.49 points, giving low/medium/high ranges $[39.55, 40.94]/[45.64, 46.49]/[47.36, 47.85]\%$, separating high effort. These are empirical-distribution identification ranges, not confidence intervals: finite-response uncertainty is not eliminated. The choice agrees with direct evaluation and pass@1; no efficiency advantage is claimed. Overlapping ranges leave an admissible set, not a forced winner. The bounds retain question-level heterogeneity.

3 Experimental Protocol

Public response banks test configuration rankings and identification bounds; controlled training tests residuals after RLVR; fixed-pool verifier experiments test selection and execution. Their sampling and inference protocols differ.

Data and model. The primary study starts from Qwen2.5-Math-1.5B (Yang et al., 2024). We sample 10,000 unique MATH-plus prompts (TIGER-Lab, 2024), remove exact normalized MATH-500 question matches (zero remain) from the MATH benchmark (Hendrycks et al., 2021), and regenerate verifier labels with a balanced boxed-answer parser. Of four teacher trajectories per prompt, the first verified-correct trajectory exists for 4,180 prompts and defines the RFT training set. Training, development, MATH-500, and AIME questions are disjoint by construction.

Re-parsing changes 2,643 labels and exposes 743 duplicate rows in the preliminary data. Only regenerated unique data enter training (Table 5).

Canonical training. We use TRL's full-parameter GRPOTrainer for 400 optimizer steps, group size 16, two policy iterations per rollout batch, clipped importance ratios, KL coefficient 0.04, and the standard GRPO loss. Seeds are 42, 1, and 2. Every rollout contributes and importance-ratio clipping is active. SAPO uses the identical loop and resource budget, replacing only its reward signal. The appendix records checkpoint retention, software versions, hardware, and generation parameters.

Evaluation and uncertainty. MATH-500 uses 256 responses per problem at temperature 0.7. We report exact pass@1 and pass@256; the original scaling curves use 1,000 deterministic subsets at intermediate K. The eight-system ranking audit instead uses exact oracle expectations, exact plurality at $K = 1, 2, 4$, 200 shared subsets at other intermediate budgets, and saved order at $K = 256$. Model comparisons use problem-paired bootstrap intervals, while training variation is summarized across the three independently trained seeds and with a hierarchical problem-within-seed bootstrap. AIME 2024 is a 30-problem stress test, not a high-powered confirmatory benchmark (Mathematical Association of America, 2024).

Cross-benchmark replication. Before generation, we freeze a seed-2027 sample of 500 from the 581 eligible OlympiadBench English open-ended competition-math items (He et al., 2024). Eligibility requires a text-only, single-answer, automatically gradable Algebra, Geometry, Combinatorics, or Number Theory problem; proof, physics, multimodal, and multiple-answer items are excluded. Base and all three canonical GRPO seeds use K=64, temperature 0.7, and one shared decoding protocol. Correctness uses the pinned official symbolic judge; question IDs, data fingerprint, code revision, and the zero exact overlap with the 10,000 training prompts are recorded before any accuracy is computed.

RFT matching. RFT performs full-parameter SFT for 800 steps on the 4,180 verified teacher trajectories. Every 100 steps, checkpoints are evaluated on 500 audited development problems disjoint from training, MATH-500, and AIME. The fixed rule selects the closest pass@1 to GRPO, with earlier steps breaking ties; MATH-500 remains locked unless the difference is within 1.5 points. All checkpoints and outcomes are retained; selection precedes test access.

Practical selectors and boundaries. At $K = 64$, Qwen2.5-Math-PRM-7B (Zhang et al., 2025) supplies direct Best-of-N and score-weighted consensus. Product aggregation is primary; terminal and minimum scores are fixed sensitivities. The software stack must reproduce a pinned official example before scoring, and ProcessBench tests step discrimination. All 36 PRM–plurality comparisons share simultaneous intervals. A separate RM-72B experiment (Yang et al., 2024) is excluded pending its official-example validation (Appendix P). A frozen 7B-Instruct model instead solves each problem once and selects its matching cluster, falling back to plurality. Neither sees test labels; exact full-pool order is preserved. Separately, released ProRL and base checkpoints (Liu et al., 2025a) are sampled with their documented long-response setting. Because that protocol differs from ours, ProRL is a boundary case, not a direct row in the primary matched-pass@1 reference comparison.

Policy-aware 7B scalar verifiers use a disjoint 400/100-question fit/tuning split; mixed-policy fits have three seeds. A generative candidate judge ranks the same top-eight structured clusters, with its prompt selected on disjoint development. Both are frozen before label-hidden test scoring. The appendix reports transfer fits, candidate-order controls, and a separately labeled format-constrained follow-up (Stephan et al., 2024).

Non-mathematical diagnostic. On 257 sanitized MBPP tasks (Austin et al., 2021), a frozen $K{=}32$ selector maximizes public tests passed; correctness uses hidden tests. Pinned DeepCoder-14B and its same-backbone initializer (Luo et al., 2025) share this contract. A cap-triggered 16K extension and its interpretation thresholds are frozen before extension outputs (appendix).

4 RESULTS

4.1 HIGH-BUDGET CONFIGURATION CHOICE AND LOW-BUDGET MECHANISM

We reanalyze the public banks of Hariri et al. (2026a), freezing the analysis before loading response-level outcomes. Four configurations cover Qwen3.6-35B-A3B and gpt-oss-20b at low/medium/high effort, each with 80 responses on 3,600 SuperGPQA and 186 math problems. All six pairs, three budgets $(32, 64, 80)$, four endpoints, and both tasks form one 144-contrast family. On SuperGPQA at $K = 80$, low effort exceeds high effort in oracle coverage by 2.86 points (simultaneous 95% CI $[0.91, 4.81]$), yet trails in plurality by 7.64 ($[5.31, 9.97]$) and reference-free BoN by 5.08 ($[2.65, 7.51]$). Thus oracle-based choice between low and high loses 7.64 plurality points. Among three effort settings, oracle selects medium but plurality selects high, losing 1.89 points ($[0.39, 3.39]$); among all four systems both select Qwen, with zero loss (Table 39). Both effort-level plurality reversals survive invalid voting and $K = 32, 64$ (Figure 1a), but none appear on math or between backbones. These comparisons share banks and differ in compute (mean completion lengths: 355/1,767/10,671 tokens for low/medium/high). Source aggregates preceded this secondary analysis (Appendix V); it is not a blinded independent replication.

Across eight systems, RFT leads oracle at $K = 2, 4$ under both conservative mappers, while plurality with abstention favors GRPO-42. Structured-semantic GRPO-42–RFT oracle differences are $-1.61/-2.59$ points and plurality differences $+4.53/+2.38$. Pointwise intervals resolve both reversals, but maximum-statistic correction over all $28 \times 9 \times 3 \times 2 = 1{,}512$ contrasts resolves neither. Six RL–RFT pairs share one reference: they are correlated exploratory examples, not independent confirmations (Appendix F).

For N saved responses with c correct and i invalid, first-occurrence ties give exactly

$$A_{\text{abstain}}(2) = \frac{c}{N} + \frac{ic}{N(N-1)}, \qquad A_{\text{invalid votes}}(2) = \frac{c}{N}. \tag{5}$$

The $+4.53$-point difference decomposes into -0.014 points in pass@1 and $+4.543$ in invalid-first/correct-second recovery. Making invalid outputs one ordinary wrong category changes the $K = 4$ difference to -6.44 points, while oracle coverage is unchanged. Thus this low-budget reversal is mediated by validity filtering. It has a different mechanism from the high-budget configuration reversals above, which survive the same invalid-vote control.

4.2 MAPPER-ROBUST RESIDUAL, NO STABLE TRAINING DIRECTION

Table 1: MATH-500 residual at $K{=}256$ (percentage points). GRPO-42 and the three-seed mean are shown separately; Δ_{42} compares GRPO-42 with Base by paired problem bootstrap.

Mapper	Base	GRPO-42	Mean	Δ_{42} (95% CI)
Released fallback	27.2	23.8	23.27	-3.4 $[-5.8, -1.0]$
Complete box	20.0	20.8	19.60	$+0.8$ $[-1.2, 2.8]$
Structured semantic	19.4	19.8	18.87	$+0.4$ $[-1.4, 2.2]$

Complete-box and structured-semantic voting leave 18.9–20.0-point residuals (Table 1). Neither mapper shows a distinguishable GRPO–Base change. The released fallback instead gives a -3.4-point seed-42 contrast because every unboxed last-line string becomes a category; it is a compatibility endpoint, not the basis for a training-direction claim. Across all three GRPO seeds, the complete-box gap is 19.60 ± 1.20 and the structured-semantic gap is 18.87 ± 1.01 points. Thus the residual is mapper-robust while its training-associated direction is not.

Table 2: Two 7B LoRA architecture boundaries on MATH-500 under the structured-semantic mapper (%). GRPO is mean±SD over three seeds; comparisons are within backbone because prompt templates and stable learning rates differ.

Backbone	Update	p@1	p@256	Plur.	Gap
Qwen-7B	Base	40.96	92.6	77.8	14.8
	GRPO	64.81±2.44	91.80±.20	77.87±.61	13.93±.76
Mistral-7B	Base	11.43	76.0	27.2	48.8
	GRPO	11.95±.12	76.33±.61	25.80±.60	50.53±.70

Table 3: Pre-frozen OlympiadBench-Text500 replication at $K=64$ (%). Correctness uses the pinned official symbolic judge; GRPO is mean±sample SD over the same three canonical training seeds.

Policy	pass@1	Oracle	Plurality	Gap ↓
Base	20.01	65.0	38.8	26.2
GRPO	21.78±0.38	65.93±1.14	40.20±1.06	25.73±1.50

For completeness, under the released endpoint GRPO raises pass@1 from 27.64% to $31.73\pm0.40\%$ and plurality@256 from 64.8% to $68.40\pm0.53\%$. Pass@256 remains $91.67\pm0.42\%$, and the remaining-error-normalized gain change is nonsignificant (seed 42: $+1.92$ points, CI $[-0.87, 4.80]$, $p=0.192$). We therefore find no evidence of an intrinsic RLVR scaling penalty. Full per-seed endpoints, winner disagreements, and symbolic sensitivity appear in the appendix.

Table 2 separates scale from architecture without relying on fallback parsing. Qwen-7B retains a 13.4–14.8-point residual; its GRPO–Base gap change is -0.87 points (hierarchical 95% CI $[-3.00, 1.27]$). On the independently pretrained Mistral family (Jiang et al., 2023), the residual is 48.8–51.2 points and the change is $+1.73$ ($[-1.60, 5.00]$). Complete-box voting gives the same conclusion (Qwen 14.0–16.2; Mistral 47.6–49.4). Thus the residual transfers across families, whereas its training-associated direction again does not.

The appendix's endpoint transition audit localizes the seed-42 improvement: 22 Base misses become selected while 5 selected problems become missed, but aggregate oracle coverage is unchanged. This is a redistribution of categorical mass among already sampled answers, which pass@K alone conceals.

Table 3 replicates the residual beyond MATH-500. Complete-box and structured-semantic voting retain mean GRPO gaps of 24.93 and 25.00 points; treating length-stopped generations as abstentions retains 25.13 points. The seed-42 gap contrast is $+0.4$ points (paired 95% CI $[-2.8, 3.6]$), so this benchmark again supports a residual, not a stable training-direction claim. A disjoint 32/32 categorical estimate predicts the held-out gap within 1.1–2.0 points across all four policies. The 30-problem AIME result remains a low-powered negative stress test and is reported in full in the appendix.

A positive residual also remains after code-specialized RL. On 257 MBPP tasks at $K=32$, initializer oracle/selected accuracies are 96.11/91.44%, leaving a 4.67-point residual (95% CI $[2.33, 7.39]$); its same-backbone DeepCoder-14B checkpoint retains 95.33/93.00% and 2.33 points ($[0.78, 4.28]$). The paired gap change, -2.33 points ($[-5.06, 0.39]$), and other endpoint changes are inconclusive. A 5.19% maximum cap rate gives the frozen "qualified," not "clean," transfer label (appendix).

4.3 The answer distribution predicts a disjoint sample pool

Under the released fallback mapper, GRPO-42 reduces normalized categorical entropy from 0.607 to 0.561 and effective support from 47.5 to 38.6 (Figure 4). This is not automatically harmful: mean correct mass rises from 27.64% to 31.95%, the modal margin rises from 15.3 to 19.1 points, and the strictly recoverable fraction rises from 64.4% to 68.0%. This decomposition explains why the absolute gap improves despite concentration. It also shows why entropy alone is not an explanatory variable: one must know where the probability mass moves.

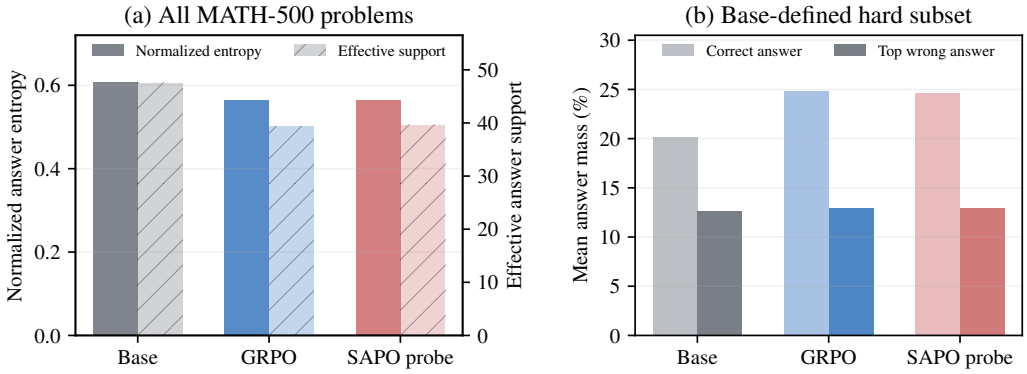

Figure 4: Categorical measurements from 256 responses per MATH-500 problem, using the released fallback mapper and GRPO-42. Entropy and effective support decrease while correct mass increases. On the Base-defined hard solvable subset, correct mass exceeds top-wrong mass but a substantial wrong mode remains.

Table 4: Structured-semantic selector spectrum on the first 64 saved responses per MATH-500 problem (%). "Work" is additional 7B inference beyond the saved pool; token counts sum mean input and output and are not FLOP estimates.

Selector	Base	GRPO-42	Extra work
Plurality	69.2	72.0	none
Candidate judge	68.2	70.4	$\sim$739 tok.
Adapted verifier	69.13$\pm$1.10	70.53$\pm$1.63	8 scalar scores
Frozen PRM-WSC	75.4	75.8	64 process scores
Frozen Gen-7B	78.4	79.0	$\sim$755 tok.
Oracle coverage	88.0	88.8	ceiling

Using the same mapper and seed, we test disjoint pools of 128 responses each. Base's plug-in predicted gap is 24.23% and Pool B observes 25.04%, an error of -0.81 points. GRPO predicts 21.27% and observes 21.65%, an error of -0.38 points. Across 100 random partitions, prediction-minus-observation split-sensitivity ranges are $[-3.04, 1.59]$ and $[-2.39, 1.85]$ points. The problem-level mean absolute errors, 12.1 and 11.1 points, are much larger than the aggregate errors. These errors limit the account's use for individual problems, despite its accurate aggregate predictions. The appendix reports near-tie bins, threshold sensitivity, and disjoint-pool calibration (Figure 5).

A fixed-target sample-complexity audit holds Pool B at 64 responses while growing a disjoint Pool A from 16 to 128. On Qwen-1.5B, aggregate absolute error stays near one point, but problem-level MAE falls from 19.2 to 12.4 points for Base and from 16.8 to 10.8 for GRPO-42; the highest-support quartile remains hardest. Mistral exposes the estimator's finite-sample limit more sharply: mean underprediction contracts from 13.2/13.8 points at 16 samples to 1.7/2.2 at 128 for Base/GRPO-42. Same-budget direct gap estimates have comparable aggregate error and lower problem MAE, whereas the plug-in has lower Brier loss (appendix). Its value is a distributional account across budgets, not a uniform prediction or sample-efficiency advantage.

4.4 Selector accuracy and query efficiency

Released ProRL improves every accuracy endpoint despite lower entropy and normalized gain (appendix).

Strictly held-out mixed adaptation does not beat structured plurality (Table 4): changes are -0.07 points for Base (hierarchical 95% CI $[-2.53, 2.13]$) and -1.47 for GRPO-42 ($[-3.93, 0.67]$). Frozen PRM product-weighted voting instead gains 6.2/3.8 points, with 36-contrast simultaneous intervals $[2.34, 10.06]/[0.48, 7.12]$. Official-example and ProcessBench checks validate the corrected interface (appendix). The remaining oracle residuals are 12.6/13.0 points.

Frozen Gen-7B gains 9.2/7.0 points over plurality and remains 9.6/9.8 below oracle. Yet its standalone accuracy is 77.8%: the pool adds only 0.6/1.2 points, with simultaneous intervals spanning zero. Its large gain mainly reflects the stronger solver; PRM-WSC demonstrates recovery by scoring the existing candidates.

Decision-preserving execution. A fixed-pool application of bound-based aggregation (Fagin et al., 2002) avoids unnecessary verifier calls. For weights in $[0, 1]$, a category's partial sum L_a and unqueried count n_a bound its full score by $[L_a, L_a + n_a]$. We query fixed blocks until the observations determine the full weighted-vote winner, including ties and zero-score fallback (Appendix X.1). On 64 questions per policy, this preserves full fixed-batch decisions and reduces verifier-stage time by 51–70%; candidate generation is unchanged. The guarantee concerns a fixed scoring rule, not whether its winner is correct. Full and lazy executions retain the same batch context; scores need not agree across different partitions.

5 Related Work

Coverage–selection separation appears in protocol-level evaluation (Hariri et al., 2026a), modal-limit analysis (Bay & Yearick, 2026), and open-ended generation (Romano et al., 2026). Bayesian evaluation and dense ranking quantify uncertainty in success rates (Hariri et al., 2026c;b). Our question is cross-system identification: even exact oracle curves cannot determine deployed model choice. Multiclass voting bounds (Aeeneh et al., 2023) describe error under a specified probability model. Here we solve a different extremal problem: the sharp accuracy range over all wrong-answer distributions consistent with retained summary information. We then test system choices with simultaneous uncertainty.

Test-time scaling includes self-consistency, adaptive compute allocation, budget forcing, and test-time updates (Wang et al., 2023; Damani et al., 2025; Muennighoff et al., 2025; Zuo et al., 2025). Best-of-Majority gives within-generator top-k minimax guarantees (Di et al., 2026); Weaver ensembles weak verifiers to narrow the generation–verification gap (Saad-Falcon et al., 2025). Verifiers strengthen selection (Cobbe et al., 2021; Lightman et al., 2024; Luo et al., 2024; Setlur et al., 2024; Zhang et al., 2024), but verifier reliability and generalization remain limiting (Zheng et al., 2024; Stroebl et al., 2024). Our appendix also reports the heuristic SAPO intervention and its relation to multi-sample leave-one-out optimization (Tang et al., 2025).

6 Discussion and Conclusions

A retrospective question-split pilot selects systems on 1,800 SuperGPQA questions before evaluating the other 1,800. At $K = 64$, direct plurality evaluation selects high effort rather than oracle's medium, gaining 2.32 points (48-contrast simultaneous CI $[0.91, 3.73]$). Pass@1 and categorical prediction make the same choice; all four rules choose Qwen in this block when included. This supports candidate-set-specific decision cost, not superiority over pass@1 or prospective confirmation (Appendix W).

At $K = 64$, the 1.5B pool contains 70–76k re-encoded output tokens per problem; PRM adds 77–83k input tokens. Standalone 7B uses 656 output-text tokens at 77.8% accuracy. These counts are not FLOPs or latency and establish no compute-optimal pipeline (Appendix P.6).

Full-parameter controlled training is limited to 1.5B mathematical reasoning. The 14B code-RL boundary and RFT are not compute-matched; the former's 5.19% cap rate precludes a clean label. Pretraining contamination remains possible, and answer matching does not certify traces (Wen et al., 2025). PRM/ProRL protocols differ; claims remain indexed by model, decoding, mapper, selector, and budget.

Roughly one third of GRPO responses reach or approach the 2,048-token cap. A paired 100-problem extension to 3,072 tokens retains 19–21-point structured-semantic residuals, with 30–36% still capped (Appendix H.1). This supports persistence, not length invariance.

Oracle curves constrain coverage, not deployed rankings. Identification bounds expose the missing categorical information; reliable system choice still requires measuring the intended selector.

AI USE STATEMENT

Generative AI tools were used to provide feedback on research methodology and experimental design, assist with theoretical formulation, proofs, implementation, data preparation, and result interpretation, and support language editing. The authors retained full control over all scientific and experimental decisions, independently validated the results and conclusions, and take full responsibility for the final work.

ETHICS STATEMENT

This work uses public models and benchmark datasets and involves no human subjects or private data. Its main deployment implication is cautionary: oracle coverage should not be reported as selected system accuracy. We release aggregate analyses and an anonymized artifact while excluding machine-local and identity-bearing metadata.

REPRODUCIBILITY STATEMENT

Section 3 specifies the primary training, sampling, selector, and uncertainty protocol. The appendix provides complete proofs, model revisions, data and parser audits, per-seed results, selector sensitivities, compute details, and an artifact map. The anonymous code/data archive contains the analysis and evaluation programs, tests, immutable manifests, and machine-readable evidence used for every reported result.

REFERENCES

Sina Aeeneh, Nikola Zlatanov, and Jiangshan Yu. New bounds on the accuracy of majority voting for multi-class classification. *arXiv preprint*, 2023. URL https://arxiv.org/abs/2309.09564. arXiv:2309.09564.

AI-MO. AIMO validation AIME. Hugging Face dataset, 2024. URL https://huggingface.co/datasets/AI-MO/aimo-validation-aime. Accessed: 2026-07-22; filtered to the 30 AIME 2024 I/II problems.

Siddharth Aphale and Kelly Liu. SFT overtraining predicts rank inversion via entropy collapse under RLVR, 2026. arXiv:2606.18487.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc V. Le, and Charles Sutton. Program synthesis with large language models, 2021. arXiv:2108.07732.

Yong Yi Bay and Kathleen A. Yearick. When more sampling hurts: The modal ceiling and correlation ceiling of test-time scaling, 2026. URL https://arxiv.org/abs/2606.28661. arXiv:2606.28661.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling, 2024. arXiv:2407.21787.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, et al. Evaluating large language models trained on code, 2021. URL https://arxiv.org/abs/2107.03374. arXiv:2107.03374.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. arXiv:2110.14168.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Yuchen Zhang, Jiacheng Chen, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, et al. Process reinforcement through implicit rewards, 2025a. arXiv:2502.01456.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, et al. The entropy mechanism of reinforcement learning for reasoning language models, 2025b. arXiv:2505.22617.

Mehul Damani, Idan Shenfeld, Andi Peng, Andreea Bobu, and Jacob Andreas. Learning how hard to think: Input-adaptive allocation of LM computation. In *International Conference on Learning Representations*, 2025.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning, 2025. arXiv:2501.12948.

Qiwei Di, Kaixuan Ji, Xuheng Li, Heyang Zhao, and Quanquan Gu. Best-of-majority: Minimax-optimal strategy for pass@k inference scaling. In *International Conference on Learning Representations*, 2026. arXiv:2510.03199.

Ron Fagin, Amnon Lotem, and Moni Naor. Optimal aggregation algorithms for middleware. *arXiv preprint*, 2002. URL https://arxiv.org/abs/cs/0204046. arXiv:cs/0204046.

Wei Fu, Jiaxuan Gao, Xujie Shen, Chen Zhu, Zhiyu Mei, Chuyi He, Shusheng Xu, Guo Wei, Jun Mei, Jiashu Wang, et al. AReaL: A large-scale asynchronous reinforcement learning system for language reasoning, 2025. arXiv:2505.24298.

Mohsen Hariri, Weicong Chen, Nahal Shahini, Vikash Singh, Kai Ye, Amirhossein Samandar, Debargha Ganguly, Sreehari Sankar, Yanyan Zhang, Shouren Wang, Jerry Peng, Biyao Zhang, Michael Hinczewski, and Vipin Chaudhary. Test-time scaling in reasoning LLMs: Inference regimes, evaluation, and reproducibility, 2026a. URL https://arxiv.org/abs/2608.04001. arXiv:2608.04001.

Mohsen Hariri, Michael Hinczewski, Jing Ma, and Vipin Chaudhary. Ranking reasoning LLMs under test-time scaling, 2026b. URL https://arxiv.org/abs/2603.10960. arXiv:2603.10960.

Mohsen Hariri, Amirhossein Samandar, Michael Hinczewski, and Vipin Chaudhary. Don't pass@k: A Bayesian framework for large language model evaluation. In *International Conference on Learning Representations*, 2026c. URL https://arxiv.org/abs/2510.04265.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3828–3850, 2024. doi: 10.18653/v1/2024.acl-long.211. URL https://aclanthology.org/2024.acl-long.211/.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. arXiv:2103.03874.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. Open-Reasoner-Zero: An open source approach to scaling up reinforcement learning on the base model, 2025. arXiv:2503.24290.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7B, 2023. arXiv:2310.06825.

Yuxian Jiang, Yafu Li, Guanxu Chen, Dongrui Liu, Yu Cheng, and Jing Shao. Rethinking entropy regularization in large reasoning models, 2025. arXiv:2509.25133.

Leslie Kish. *Survey Sampling*. John Wiley & Sons, New York, 1965.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024.

Zhimin Lin, Yixin Ji, Jinpeng Li, Yu Luo, Dong Li, Junhua Fang, Juntao Li, and Min Zhang. When to vote, when to rewrite: Disagreement-guided strategy routing for test-time scaling, 2026. arXiv:2604.26644.

Mingjie Liu, Shizhe Diao, Ximing Lu, Jian Hu, Xin Dong, Yejin Choi, Jan Kautz, and Yi Dong. ProRL: Prolonged reinforcement learning expands reasoning boundaries in large language models. In *Advances in Neural Information Processing Systems*, 2025a.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding R1-Zero-like training: A critical perspective, 2025b. arXiv:2503.20783.

Anamika Lochab, Bolian Li, and Ruqi Zhang. Uniform-correct policy optimization: Breaking RLVR's indifference to diversity, 2026. arXiv:2605.00365.

Liangchen Luo, Yinxiao Liu, Rosanne Liu, Samrat Phatale, Meiqi Guo, Harsh Lara, Yunxuan Li, Lei Shu, Yun Zhu, Lei Meng, et al. Improve mathematical reasoning in language models by automated process supervision, 2024. arXiv:2406.06592.

Michael Luo, Sijun Tan, Roy Huang, Ameen Patel, Alpay Ariyak, Qingyang Wu, Xiaoxiang Shi, Rachel Xin, Colin Cai, Maurice Weber, Ce Zhang, Li Erran Li, Raluca Ada Popa, and Ion Stoica. DeepCoder: A fully open-source 14b coder at o3-mini level. Hugging Face model card, 2025. URL https://huggingface.co/agentica-org/DeepCoder-14B-Preview. Accessed: 2026-08-31.

Mathematical Association of America. AIME 2024: American invitational mathematics examination. Mathematical Association of America, 2024. URL https://maa.org/maa-invitational-competitions/. Accessed: 2026-07-22.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling, 2025. arXiv:2501.19393.

Adam Ousherovitch and Ambuj Tewari. Compute aligned training: Optimizing for test time inference. *arXiv preprint arXiv:2604.24957*, 2026. URL https://arxiv.org/abs/2604.24957. arXiv:2604.24957.

Yossef Rinott. Multivariate majorization and rearrangement inequalities with some applications to probability and statistics. *Israel Journal of Mathematics*, 15:60–77, 1973. doi: 10.1007/BF02771774. URL https://doi.org/10.1007/BF02771774.

Davide Romano, Kanak Raj, Jerrod Parker, and Daniele Giofrè. Test-time scaling in the wild: Why exploitation, not exploration, is the bottleneck, 2026. URL https://arxiv.org/abs/2608.18931. arXiv:2608.18931.

Jon Saad-Falcon, E. Kelly Buchanan, Mayee F. Chen, Tzu-Heng Huang, Brendan McLaughlin, Tanvir Bhathal, Shang Zhu, Ben Athiwaratkun, Frederic Sala, Scott Linderman, Azalia Mirhoseini, and Christopher Ré. WEAVER: Shrinking the generation-verification gap with weak verifiers. In *Advances in Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=dRjt4vlYVQ.

Kusha Sareen, Morgane M. Moss, Alessandro Sordoni, Rishabh Agarwal, and Arian Hosseini. Putting the value back in RL: Better test-time scaling by unifying LLM reasoners with verifiers, 2025. arXiv:2505.04842.

Amrith Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. Rewarding progress: Scaling automated process verifiers for LLM reasoning, 2024. arXiv:2410.08146.

Amrith Setlur, Nived Rajaraman, Sergey Levine, and Aviral Kumar. Scaling test-time compute without verification or RL is suboptimal, 2025. arXiv:2502.12118.

Nisarg Shah and Ziqi Yu. The art of calling the winner by asking just enough questions. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pp. 6149–6164, 2026. URL https://proceedings.mlr.press/v337/shah26a.html.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models, 2024. arXiv:2402.03300.

Lloyd S. Shapley. A value for n-person games. *Contributions to the Theory of Games*, 2:307–317, 1953.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *International Conference on Learning Representations*, 2025. arXiv:2408.03314.

Andreas Stephan, Dawei Zhu, Matthias Aßenmacher, Xiaoyu Shen, and Benjamin Roth. From calculation to adjudication: Examining LLM judges on mathematical reasoning tasks, 2024. arXiv:2409.04168.

Benedikt Stroebl, Sayash Kapoor, and Arvind Narayanan. The limits of inference scaling through resampling, 2024. arXiv:2411.17501.

Yunhao Tang, Kunhao Zheng, Gabriel Synnaeve, and Remi Munos. Optimizing language models for inference time objectives using reinforcement learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 59066–59085, 2025. URL https://proceedings.mlr.press/v267/tang25o.html.

TIGER-Lab. MATH-plus. Hugging Face dataset, revision 9a6d00d, 2024. URL https://huggingface.co/datasets/TIGER-Lab/MATH-plus. Accessed: 2026-07-22.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process- and outcome-based feedback, 2022. arXiv:2211.14275.

Ante Wang, Linfeng Song, Ye Tian, Baolin Peng, Lifeng Jin, Haitao Mi, Jinsong Su, and Dong Yu. Self-consistency boosts calibration for math reasoning, 2024a. arXiv:2403.09849.

Peiyi Wang, Lei Li, Zhihong Shao, R. X. Xu, Damai Dai, Yifei Li, Deli Chen, Y. Wu, and Zhifang Sui. Math-Shepherd: Verify and reinforce LLMs step-by-step without human annotations, 2024b. arXiv:2312.08935.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, et al. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for LLM reasoning, 2025. arXiv:2506.01939.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023.

Xumeng Wen, Zihan Liu, Shun Zheng, Shengyu Ye, Zhirong Wu, Yang Wang, Zhijian Xu, Xiao Liang, Junjie Li, Ziming Miao, Jiang Bian, and Mao Yang. Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base LLMs, 2025. arXiv:2506.14245.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. Qwen2.5-Math technical report: Toward mathematical expert model via self-improvement, 2024. arXiv:2409.12122.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, Lingjun Liu, Xin Liu, et al. DAPO: An open-source LLM reinforcement learning system at scale. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Lifan Yuan, Wendi Li, Huayu Chen, Ganqu Cui, Ning Ding, Kaiyan Zhang, Bowen Zhou, Zhiyuan Liu, and Hao Peng. Free process rewards without process labels, 2024. arXiv:2412.01981.

Yang Yue, 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? In *Advances in Neural Information Processing Systems*, 2025.

Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun Ma, and Junxian He. SimpleRL-Zoo: Investigating and taming zero reinforcement learning for open base models in the wild, 2025. arXiv:2503.18892.

Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. Generative verifiers: Reward modeling as next-token prediction, 2024. arXiv:2408.15240.

Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. The lessons of developing process reward models in mathematical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 10495–10516. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.547. URL https://aclanthology.org/2025.findings-acl.547/.

Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. ProcessBench: Identifying process errors in mathematical reasoning, 2024. arXiv:2412.06559.

Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Yuchen Zhang, Xinwei Long, Ermo Hua, Biqing Qi, Youbang Sun, Zhiyuan Ma, Lifan Yuan, Ning Ding, and Bowen Zhou. TTRL: Test-time reinforcement learning. In *Advances in Neural Information Processing Systems*, 2025.

A Appendix overview

This appendix supplies proofs, exact protocols, data and parser audits, per-K results, disjoint-pool calibration, an OlympiadBench-Text500 replication, matched rejection-SFT selection, released-ProRL analysis, frozen practical-selector evaluations, SAPO implementation details, negative results, Qwen and Mistral 7B LoRA boundaries, compute accounting, and an artifact map. Corrected evidence is kept separate from noncomparable preliminary runs throughout. Appendix V adds a separately frozen secondary analysis of public math and SuperGPQA banks, with high-budget configuration comparisons and simultaneous inference.

B Scope and Status of Evidence

The oracle–plurality residual remains large at high K under two conservative answer mappers and the released compatibility endpoint, while only the latter yields an apparent GRPO-associated gap reduction. A categorical account estimated from disjoint responses predicts the aggregate residual. A strictly held-out adapted verifier does not improve conservative plurality, whereas a frozen independent 7B solver-selector substantially outperforms it but still leaves a 9.6–10.2-point oracle residual across mappers.

The three-seed Mistral-7B comparison retains a 47.6–51.2-point residual across conservative mappers, while its GRPO-associated change is inconclusive. A frozen MBPP diagnostic retains a positive residual under a same-backbone 14B code-RL pair when correctness is defined by held-out executable tests. Canonical GRPO improves pass@1, but neither conservative mapper supports a stable high-K training direction. These results do not show that RLVR universally creates a scaling penalty. The change in remaining-error-normalized gain is nonsignificant in the completed paired run, and released ProRL provides an explicit boundary case.

The public-bank extension resolves high-budget gpt-oss effort-setting reversals on SuperGPQA under both plurality validity rules. It does not establish a cross-backbone reversal, a math reversal, an equal-compute advantage, or a causal training effect. Its comparison family is separate from the original eight-system training audit, whose low-budget crossings remain exploratory.

All headline experiments use the corrected balanced answer parser and a canonical TRL implementation. Clean-data, three-seed Qwen and Mistral 7B LoRA experiments are reported strictly as boundaries. Preliminary 7B experiments used corrupted nested- brace labels and a modified one-inner-step training loop. Their numerical results are not reused as replication evidence, not pooled with corrected runs, and not reproduced in the paper tables. We retain only cryptographic manifests and audit summaries so that this exclusion itself is reproducible.

C Definitions and Exact Identities

C.1 Finite-sample metrics

For a problem x with correct normalized answer y^*, let $a_i = \mathrm{ans}(o_i)$ for i.i.d. responses $o_i \sim \pi_\theta(\cdot \mid x)$. With a fixed deterministic first-occurrence tie rule,

$$\mathrm{pass}_x\,@K = \mathbb{1}[\exists i \le K : a_i = y^*], \tag{6}$$

$$\mathrm{maj}_x\,@K = \mathbb{1}[\mathrm{plurality}(a_{1:K}) = y^*], \tag{7}$$

$$\mathrm{gap}_x\,@K = \mathrm{pass}_x\,@K - \mathrm{maj}_x\,@K. \tag{8}$$

The reported dataset metrics average over problems. At $K = 1$ the oracle, plurality, Best-of-1, and weighted-self-consistency selectors are identical. A unit test checks exact rather than Monte Carlo equality.

Conditional on categorical counts, i.i.d. draws are uniform over their distinct orderings. Categories tied at the largest eligible count have equal multiplicity, so each is equally likely to occur first among those categories. Their conditional winning probabilities are therefore $1/|M(n)|$, giving $\tau_*(n)$ in Eq. 4. This integrates first-occurrence ties exactly; it does not replace the realized tie rule on a saved ordered pool. Exhaustive enumeration through $K = 6$ checks both invalid-vote and abstention contracts.

We also summarize each problem's categorical answer distribution $\hat{p}_j$ directly. Normalized entropy is $H_{\mathrm{norm}} = H(\hat{p}_j)/\log K$ and effective support is $e^{H(\hat{p}_j)}$, the entropy-based analogue of an effective sample size (Kish, 1965): it reports how many answers the distribution behaves as if it were spread over, rather than how many distinct strings appeared. Neither quantity identifies selection outcomes by itself, which is why we always report them beside correct mass, top-wrong mass, and the modal margin.

The normalized marginal gain is

$$G_K = \frac{A_{\mathrm{maj}}@K - A_{\mathrm{maj}}@1}{1 - A_{\mathrm{maj}}@1}. \tag{9}$$

It measures the fraction of the error remaining after one sample that plurality removes with K samples. It is not a causal effect and does not fully remove problem-composition changes induced by training.

C.2 Raw LOO policy-gradient identity

For a group $G = (o_1, \ldots, o_n)$, define selected utility $U(G) = \mathbb{1}[\mathrm{plurality}(G) = y^*]$ and $c_i = U(G) - U(G \setminus i)$.

Proposition 3 (Exact raw-credit identity)**.** *Let $J(\theta) = \mathbb{E}_{x,G\sim\pi_\theta^n}[U(G)]$, assuming the usual conditions for interchanging gradient and expectation. Then*

$$\nabla_\theta J(\theta) = \mathbb{E}\left[\sum_{i=1}^{n} c_i \nabla_\theta \log \pi_\theta(o_i \mid x)\right]. \tag{10}$$

Proof. The score-function identity gives

$$\nabla_\theta J = \mathbb{E}\left[U(G)\sum_i \nabla_\theta \log \pi_\theta(o_i \mid x)\right]. \tag{11}$$

Conditional on x and $G \setminus i$, the baseline $U(G \setminus i)$ is independent of o_i. Therefore

$$\mathbb{E}[U(G \setminus i)\nabla_\theta \log \pi_\theta(o_i \mid x)] = 0. \tag{12}$$

Subtracting this valid baseline separately for every score term yields Eq. 10. There is no $O(n^{-1/2})$ approximation. □

This is the standard multi-sample leave-one-out policy-gradient identity, including the majority-vote specialization of Tang et al. (2025), not a new estimator. The credit c_i is the leave-one-out marginal contribution of rollout o_i to the group's selected utility, the same primitive that underlies marginal-contribution credit assignment in cooperative games (Shapley, 1953); here a single leave-one-out term suffices because the baseline only has to be conditionally independent of o_i, not permutation-symmetric.

This proposition applies to raw c_i and sequence log probability. The actual probe separately standardizes binary verifier rewards and LOO credits, mixes them, and optimizes with clipped ratios, two inner iterations, token masks, and KL regularization. Those operations are not invariant transformations of the gradient. Accordingly, implemented SAPO is a heuristic diagnostic inspired by the identity, not an unbiased estimator of ∇J.

D Plurality Limit with Countable Answers and Ties

First assume all categories vote. Fix a problem and write p_* for correct-answer probability. Because the answer space is countable, let $p_w = \max_{a \neq y^*} p(a)$ and $\Delta = p_* - p_w$. A maximum exists whenever the supremum is positive: otherwise infinitely many atoms bounded away from zero would have divergent total mass.

D.1 Oracle-equivalent and oracle-dominant constructions

For a fixed problem, the complete oracle curve is in bijection with the single scalar p_*: $O(K) = 1-(1-p_*)^K$ and $O(1) = p_*$. It therefore contains no information about how the remaining $1-p_*$ mass is allocated among wrong answers. The following construction varies exactly that unidentified object.

Proposition 1 follows from two explicit finite distributions. Fix $q \in (0, 1/2)$, $0 \leq \epsilon < 1/2 - q$, and $m > (1-q)/q$. Distribution P assigns mass $q+\epsilon$ to the correct answer and $1-q-\epsilon$ to one wrong answer. Distribution Q assigns mass q to the correct answer and divides $1-q$ equally among m wrong answers. Their oracle success probabilities are

$$O_P(K) = 1-(1-q-\epsilon)^K, \qquad O_Q(K) = 1-(1-q)^K. \tag{13}$$

For $\epsilon = 0$ the complete oracle curves are identical. For every $\epsilon > 0$, $O_P(K) > O_Q(K)$ at all finite $K \geq 1$, so oracle-based choice always returns P.

Yet $1-q-\epsilon > q+\epsilon$, making the wrong answer the unique mode under P, whereas $(1-q)/m < q$, making the correct answer the unique mode under Q. The strong law sends plurality accuracy to zero for P and to one for Q; strict margins make this independent of finite-sample ties. For all sufficiently large K, consequently,

$$\mathcal{R}_{\mathrm{maj}}(K) = A^Q_{\mathrm{maj}}(K) - A^P_{\mathrm{maj}}(K) \longrightarrow 1. \tag{14}$$

Thus oracle choice can incur the maximum possible asymptotic deployment regret even when its winner strictly dominates in coverage at every budget. Equivalently, any model-selection rule required to choose a strict oracle dominator has worst-case asymptotic plurality regret one: an oracle-only recalibration cannot provide a nontrivial guarantee without information about the deployed selector.

The equal-curve case gives a separate indistinguishability lower bound. Give the two candidate systems arbitrary labels 1 and 2. A rule whose input is only their complete oracle curves sees the same input whether (P, Q) is assigned to $(1, 2)$ or to $(2, 1)$. A deterministic rule therefore makes the same labeled choice in both worlds. Even if that choice varies with K, one label is selected infinitely often, giving limsup regret one in one world. For a randomized rule with choice probability r_K, the expected regrets in the two worlds are $r_K + o(1)$ and $1 - r_K + o(1)$; at least one world therefore has limsup expected regret at least $1/2$. Randomizing uniformly attains this two-system lower bound. The obstruction is informational rather than a poor choice of monotone transform of pass@K.

The claim is also finite-budget quantitative. Let $\delta_P = 1-2(q+\epsilon) > 0$ and $\delta_Q = q-(1-q)/m > 0$. Applying Proposition 5 to both systems gives

$$\mathcal{R}_{\mathrm{maj}}(K) \geq 1 - me^{-K\delta_Q^2/2} - e^{-K\delta_P^2/2}. \tag{15}$$

Whenever the right side is positive, the oracle-dominant system is already worse under plurality at that finite budget; the bound approaches one exponentially. This proves possibility, not prevalence. The uniform eight-system audit separately measures where resolved reversals occur in the observed policies.

D.2 Operational countable-space result

Proposition 4 (Operational plurality limit)**.** *For fixed $\epsilon, \delta > 0$, if $p_* \geq \epsilon$ and $\Delta \geq \delta$, then pass@K and plurality accuracy converge to one. If $p_* \geq \epsilon$ and $\Delta \leq -\delta$, then pass@K converges to one and plurality accuracy converges to zero. Consequently, the aggregate gap has asymptotic lower bound*

$$\liminf_{K\to\infty} \overline{\mathrm{gap}\ @K} \geq \Pr[p_* \geq \epsilon, \Delta \leq -\delta]. \tag{16}$$

Cases with $|\Delta| < \delta$, including tied population modes, are an explicit remainder.

Proof. For $p_* \geq \epsilon$, oracle failure has probability $(1-p_*)^K \leq (1-\epsilon)^K \to 0$. Suppose first that the correct answer is the unique population mode with margin at least δ. Choose a finite set F containing the correct answer and every atom whose probability is at least $p_*/3$. Choose it large enough that the total probability outside F is below $p_*/3$. By the strong law, all empirical frequencies in F and the total tail frequency converge simultaneously. Eventually the correct empirical frequency exceeds every competitor in F by a positive margin and exceeds the entire tail frequency, hence it is the empirical plurality. If an incorrect answer beats the correct one by at least δ, the strong law makes their empirical count difference linear in K, so the correct answer cannot win. Problemwise convergence and Fatou's lemma yield Eq. 16. □

If all operationally solvable problems have unique population modes and p_* is either zero or bounded below by ϵ, bounded convergence gives equality with the non-recoverable fraction as $\delta \downarrow 0$. If $\Delta = 0$, empirical top counts fluctuate; the limiting correct-selection probability depends on the number of tied modes and the selector's tie rule. We therefore neither label tied modes non-recoverable nor assert their gap tends to one. The $p_* \geq \epsilon$ condition also avoids the vacuous softmax-model notion that every finite string with positive probability is "solvable."

D.3 Finite-budget modal-margin proof

Proposition 5 (Finite-budget modal-margin bound)**.** *Suppose all categories vote and the answer support has size M. If $p_* - p_w \geq \delta > 0$, plurality error at budget K is at most $(M-1)e^{-K\delta^2/2}$. If some wrong answer satisfies $p_w - p_* \geq \delta$, plurality correctness is at most $e^{-K\delta^2/2}$, and therefore*

$$\mathbb{E}[\mathrm{gap}_x\ @K] \geq 1 - (1-p_*)^K - e^{-K\delta^2/2}. \tag{17}$$

Assume finite support of size M. For a wrong answer a, define $Z_i = \mathbb{1}[A_i = y^*] - \mathbb{1}[A_i = a] \in [-1, 1]$. If the correct modal margin is at least δ, then $\mathbb{E}Z_i = p_* - p(a) \geq \delta$. Hoeffding's inequality gives

$$\Pr[N_a(K) \geq N_*(K)] = \Pr\left[\frac{1}{K}\sum_i Z_i \leq 0\right] \leq e^{-K\delta^2/2}. \tag{18}$$

Table 5: Training-data and parser audit. "Corrected" is regenerated from clean unique questions with balanced extraction.

Audit item	Preliminary	Corrected
Training rows	10,000	10,000
Unique questions	9,257	10,000
Duplicate rows	743	0
Duplicate groups	616	0
Conflicting-label groups	158	0
Exact MATH-500 overlap	0	0
Nested-label changes	2,643	0

If plurality is incorrect, at least one of the $M-1$ competitors has count no smaller than the correct count. A union bound proves $(M-1)e^{-K\delta^2/2}$. This event also covers empirical ties even when the fixed tie rule would favor the correct answer, so the inequality is conservative.

If a wrong answer w has $p_w - p_* \geq \delta$, reverse the definition of Z_i. Plurality correctness requires $N_*(K) \geq N_w(K)$ and is therefore at most $e^{-K\delta^2/2}$. Since oracle success is exactly $1-(1-p_*)^K$, subtracting the upper bound on plurality correctness proves the finite-budget lower bound in the main paper. For a countably infinite support, the asymptotic proof above remains valid, but a finite-K union bound requires an explicit tail-mass term; we do not replace it with the number of empirically observed answers.

Abstention contract. An invalid category is a competitor only when invalid outputs vote. If invalid outputs instead abstain with probability $\rho < 1$, condition on the number of valid draws $V \sim \mathrm{Binomial}(K, 1-\rho)$. For $V = v > 0$, apply the all-voting results to the eligible distribution $\widetilde{p}(a) = p(a)/(1-\rho)$ and budget v; average over V, assigning zero accuracy at $V = 0$. Thus the wrong mode is the largest eligible wrong category, never the abstaining invalid category. In particular, with only the correct category eligible, accuracy is $1-\rho^K$, not one; omitting the zero-valid-draw event would invalidate a naive union bound. At $\rho = 1$ selected accuracy is zero. These contracts do not merge distinct eligible answers using correctness labels.

E Corrected Data and Answer Extraction

E.1 Why preliminary data cannot support comparison

The preliminary extractor used a regular expression whose capture ended at the first closing brace. For nested TeX, for example `\boxed{\frac{1}{2}}`, it returned an incomplete fragment. A brace-balanced parser changed 2,643 of 10,000 stored training labels (26.43%) and 143 of 500 development labels (28.6%). Because the verifier reward was computed from these labels, the issue affects training, not only reporting; the associated checkpoints cannot be repaired by re-evaluation.

The corrected training set contains 10,000 unique MATH-plus questions (TIGER-Lab, 2024). Question normalization lowercases and removes whitespace before the exact-overlap audit; zero MATH-500 matches remain. A stronger secondary audit that removes every non-alphanumeric character also finds zero matches. These checks catch direct or formatting-only reuse, not paraphrases or pretraining contamination. Our internal records preserve every raw response. The review artifact now also releases a seed-2027 audit subset of 50 fixed MATH-500 problems for Base and GRPO-42, including all 256 raw responses per problem and expected outputs from the fallback, complete-box, and structured-semantic mappers. Regenerating the compact records from these 25,600 responses gives exact byte-level equality. The remaining full raw pools and checkpoints are too large for the compact archive.

E.2 Validity-aware and semantic selector sensitivity

The released compatibility evaluator gives every response one normalized label: it uses the final complete `\boxed{...}` expression when available, then documented answer-line and last-line fallbacks. This is a reproducible endpoint, but repeated malformed fallback strings can form a

Table 6: Full-pool selector sensitivity on MATH-500 (%). GRPO and SAPO are three-seed means; Base and RFT are single trained policies. Invalid responses remain in the pass@1 denominator for all mappers.

Answer mapper	Policy	pass@1	pass@256	Plur.@256	Gap@256	G_{256}
Released fallback	Base	27.64	92.0	64.8	27.2	51.35
	GRPO mean	31.73	91.67	68.40	23.27	53.71
	SAPO mean	31.60	91.27	67.27	24.00	52.14
	RFT	32.01	90.4	70.6	19.8	56.76
Complete box only	Base	27.59	92.0	72.0	20.0	61.33
	GRPO mean	31.68	91.67	72.07	19.60	59.11
	SAPO mean	31.55	91.13	71.53	19.60	58.41
	RFT	31.98	90.2	70.8	19.4	57.07
Structured semantic	Base	28.03	92.4	73.0	19.4	62.49
	GRPO mean	32.19	92.07	73.20	18.87	60.47
	SAPO mean	32.07	91.67	72.87	18.80	60.05
	RFT	32.42	91.0	71.4	19.6	57.68

plurality. We therefore add two conservative sensitivity endpoints. The first lets only nonempty, complete boxed answers vote. The second also accepts explicit "answer/result is" statements and canonically merges exact rational numbers and tuples of exact rationals (e.g., $0.15 = 3/20$), but never uses an arbitrary last line. Invalid responses remain errors in the pass@1 denominator; we do not condition accuracy on valid generations. The grammar is deliberately limited and is not a general symbolic verifier.

Both alternatives preserve the central residual while yielding near-zero training contrasts. Complete-box voting leaves Base/GRPO-mean gaps of 20.0/19.6; structured semantic voting leaves 19.4/18.87. For GRPO-42 minus Base, the problem-paired gap changes are $+0.8$ points under complete-box voting (95% CI $[-1.2, 2.8]$, p=0.555) and $+0.4$ under structured semantic voting (CI $[-1.4, 2.2]$, p=0.827). For RFT minus GRPO-42 they are -1.4 (CI $[-3.8, 1.0]$, p=0.335) and -0.2 (CI $[-2.6, 2.2]$, p=1.000). Normalized gain reverses direction under the alternatives, further showing that this headroom-derived quantity is mapper-sensitive rather than a training effect.

The decomposition identifies why. Among 128,000 responses per policy, 44,895/36,621/11,450 Base/GRPO-42/RFT responses lack a complete nonempty box. The structured rule recovers 1,629/1,642/847 explicit unboxed answer statements and rejects 43,266/34,979/10,603 arbitrary fallbacks. Exact-rational canonicalization additionally recognizes 501/590/528 response-level correct equivalences and raises structured plurality from 71.6/71.4/70.8 to 73.0/72.6/71.4. At the more consequential problem level, the primary and complete-box winners differ on 46/26/3 cases. The primary winner has no boxed support in 46/25/3; it is correct in 0/1/0 cases, whereas the complete-box winner is correct in 36/16/1. Thus valid unboxed answers exist, but arbitrary fallback modes account for nearly all winner changes; semantic equivalence accounts for few of them.

As a separate semantic sensitivity, we froze candidate labels for all 75 policy-blinded winner disagreements with the pinned OlympiadBench automatic symbolic judge before opening the candidate-source key (He et al., 2024). It marks the complete-box winner equivalent to the reference in 33/46, 16/26, and 1/3 Base/GRPO-42/RFT cases, versus 0/46, 1/26, and 0/3 for the fallback winner. This automatic check supports the objective decomposition but is not human adjudication. The artifact therefore also retains randomized candidate order, hidden policy identity, and an empty predefined annotation schema for two future independent readers. Because no human labels are frozen, we treat the oracle–selector residual as mapper-robust but do not claim the ranking of training methods is mapper-invariant.

F Decision-Level System Ranking Audit

The audit includes Base, three GRPO seeds, three SAPO seeds, and matched RFT at $K \in \{1, 2, 4, 8, 16, 32, 64, 128, 256\}$ under all three mappers. Oracle expectations are exact hypergeometric probabilities over each saved 256-response pool. The revised analysis evaluates plurality

Table 7: Low-budget decision audit. Counts refer to 28 pairs per row. Pointwise requires opposite signs in both unadjusted 95% intervals; simultaneous requires opposite signs in both full-audit max-T intervals. All six pointwise crossings share the same RFT reference.

Mapper	K	Oracle / plurality leader	Pointwise	Simultaneous
Complete box	2	RFT / GRPO-42	6	0
Complete box	4	RFT / GRPO-42	6	0
Structured semantic	2	RFT / GRPO-42	6	0
Structured semantic	4	RFT / GRPO-42	6	0

Table 8: Mechanism decomposition on all 500 problems (percentages or points). V_1 is first-draw validity; $\neg V_1, V_2$ is invalid-first/valid-second probability; rescue is invalid-first/correct-second accuracy gain. A is abstaining plurality and V lets all invalid outputs vote as one wrong category. Each entry is an exact expectation over the same 256-response pools.

Mapper	Policy	V_1	$\neg V_1, V_2$	Rescue	A@2	V@2	A@4	V@4
Complete box	GRPO-42	71.39	14.49	6.94	38.83	31.89	49.47	37.71
	RFT	91.05	6.58	2.37	34.35	31.98	47.02	44.35
Structured	GRPO-42	72.67	14.14	6.77	39.17	32.40	49.93	38.68
	RFT	91.72	6.17	2.23	34.64	32.42	47.55	45.12

exactly at $K = 1, 2, 4$ by enumerating categorical occupancies under sampling without replacement and first-occurrence ties. Conditional on counts, each tied modal category has the same probability of occurring first. This supersedes the earlier 1,000-subset $K = 4$ approximation. Other intermediate budgets retain the original 200 common seeded subsets; $K = 256$ uses the exact saved order. The highlighted pair and the mechanism sensitivity are retrospective analyses; we do not retrospectively designate a confirmatory primary contrast.

Every pointwise crossing is RFT versus one of the six GRPO/SAPO policies; these correlated pairs are not independent replications. Under structured mapping at $K = 2$, RFT's 49.23% oracle coverage exceeds GRPO-42's 47.62%, while its 34.64% plurality accuracy trails 39.17%. At $K = 4$, the exact values are 64.64 versus 62.05% oracle and 47.55 versus 49.93% plurality. Complete-box mapping gives the same decisions. At $K = 8, 16$ the conservative metrics select the same system; their larger-budget crossings do not pass even pointwise intervals. No reversal anywhere in the full audit passes the simultaneous analysis below.

F.1 Exact abstention decomposition

Let N, c, i count all, correct, and invalid responses in a saved pool. At $K = 2$, a valid first response wins every possible tie. The only additional correct selections arise when the first response is invalid and the second correct, with probability $ic/[N(N-1)]$. Equation 5 follows. Under i.i.d. population sampling the corresponding identity is $A_{\text{abstain}}(2) = p_* + p_* p_{\text{invalid}}$; the finite-pool expression accounts for sampling without replacement. Two valid draws cannot test the benefit of concentrating correct versus incorrect answer categories at this budget.

For structured mapping, the GRPO-42–RFT difference in pass@1 is -0.014 points (pointwise 95% CI $[-1.30, 1.29]$); the rescue difference is $+4.543$ ($[4.06, 5.03]$), yielding the observed $+4.529$ points at $K = 2$. The complete-box decomposition is $-0.080 + 4.563 = 4.483$ points. When invalid outputs vote, these $K = 2$ plurality differences reduce exactly to the pass@1 differences. At $K = 4$, the structured difference changes from $+2.38$ ($[1.03, 3.73]$) to -6.44 ($[-8.35, -4.55]$); the complete-box counterpart changes from $+2.45$ to -6.64 points. The oracle metric is identical across these two decision rules. Thus the highlighted ranking crossings disappear under the nonabstaining rule.

We further partition the $K = 4$ change by the number of invalid responses in the sampled set. Under structured mapping, the difference between policies in gains from abstention is $0.000, 0.029, 4.041, 4.749, 0.000$ points for zero through four invalid draws. The sum, 8.819 points,

Table 9: Structured-semantic GRPO-42 minus RFT (points). The full-audit simultaneous intervals replace interpreting pointwise discoveries as confirmatory evidence. Both metrics must exclude zero in opposite directions for a resolved reversal; neither budget meets that condition.

K	Metric	Difference	Pointwise 95%	Simultaneous 95%
2	Oracle	-1.61	$[-3.09, -0.13]$	$[-4.47, 1.24]$
2	Plurality	$+4.53$	$[3.23, 5.86]$	$[1.98, 7.08]$
4	Oracle	-2.59	$[-4.02, -1.16]$	$[-5.36, 0.18]$
4	Plurality	$+2.38$	$[1.03, 3.73]$	$[-0.23, 4.99]$

reconciles -6.438 under voting with $+2.381$ under abstention. Sets with two or three invalid responses account for almost the entire change.

As a conditional diagnostic, we integrate over independent first draws from the two policies and condition on both being valid. Both-valid probability is 68.71% for structured mapping; the conditional GRPO-42–RFT selected accuracy differences are $+11.11$ points ($[9.94, 12.31]$) at $K = 2$ and $+7.14$ ($[5.97, 8.34]$) at $K = 4$. Complete-box values are 67.23%, $+11.43$ ($[10.25, 12.64]$), and $+7.33$ ($[6.17, 8.52]$), respectively. For problem j, this calculation weights the conditional comparison by $v_{G,j}v_{R,j}$, where v is first-draw validity; it averages all possible cross-policy response pairings rather than aligning arbitrary sample indices. This conditions on a policy-dependent event, changes the target population, and at $K = 4$ still permits invalid subsequent draws. It is a descriptive diagnostic, not a causal control or additional reversal replication.

F.2 Simultaneous uncertainty over the complete search family

We retain all $28 \times 9 \times 3 = 756$ pair–budget–mapper opportunities and both oracle and plurality contrasts, for 1,512 scalar comparisons. In each of 20,000 paired bootstrap draws (seed 2027), the same resampled problem IDs are used for every system, budget, and mapper. For contrast j with mean $\hat{d}_j$ and problem-level standard error $\hat{\sigma}_j$, calculate

$$T_b = \max_j \frac{|\hat{d}_j^{(b)} - \hat{d}_j|}{\hat{\sigma}_j}, \qquad I_j = \left[\hat{d}_j - c_{.95}\hat{\sigma}_j, \hat{d}_j + c_{.95}\hat{\sigma}_j\right]. \tag{19}$$

Zero-variance, identically zero contrasts contribute zero. The empirical 95th percentile of T_b is $c_{.95} = 3.862$. These centered max-T intervals provide an approximate simultaneous bootstrap correction while preserving dependencies, including the shared RFT reference. They are conditional on the saved pools and fixed intermediate-budget Monte Carlo subsets; they are not an exact finite-sample family-wise guarantee or a training-seed population inference. The exact $K = 2, 4$ decomposition has no subset Monte Carlo error.

No complete reversal survives simultaneous correction in any mapper or budget, including the released-fallback $K = 32$–128 crossings. This limits the empirical model-selection claim; it does not alter the distributional impossibility result or the separately estimated high-budget residuals. The full records and input hashes are released in `corrected_ranking_multiplicity_audit.json` and `corrected_reversal_mechanism_audit.json`. The decomposition is checked against exhaustive enumeration of ordered draws on small synthetic pools, including ties, invalid outputs, and absent correct answers.

G RFT Construction

The completed GRPO seed-42 policy generates four trajectories for each clean training prompt. We retain the first verifier-correct trajectory and no trajectory otherwise, yielding 4,180 supervised examples. Neither development nor test outcomes participate. RFT uses full-parameter SFT; the training file, selection manifest, and every checkpoint metric are included in the anonymous artifact.

Before the locked reference comparison, we audited all four question pools under both lowercased whitespace removal and the stronger alphanumeric normalization used above. A preliminary development file created before the final clean-data export duplicated 500 training questions and is

Table 10: Corrected canonical configuration. Commands, seeds, and recorded model revisions are included in the artifact.

Component	Canonical primary setting	Purpose
Backbone	Qwen2.5-Math-1.5B (Yang et al., 2024)	public math-specialized base
Optimizer steps	400	fixed across GRPO and SAPO
Learning rate	2×10^{-6}	fixed across GRPO and SAPO
Independent seeds	42, 1, 2	training-run variation
Rollout group	16	within-prompt reward normalization
Training rollout	temperature 1.0, maximum 1,024 new tokens	on-policy generation
Policy iterations	2	nontrivial importance ratios/clipping
Loss	TRL GRPO, full rollout update	no top-four filtering
KL coefficient	0.04	explicit reference control
Evaluation samples	256 per MATH-500 problem	high-K curve
Evaluation sampling	temperature 0.7, cap 2,048, vLLM seed 0	identical across rows
Subset trials	1,000, seed 2027	intermediate-K estimates

therefore excluded from every reported selection and test. Its exclusion is recorded in the audit manifest. The replacement 500-problem development set has zero pairwise overlap with the 10,000 training questions, MATH-500, and the 30 strict AIME-2024 questions under both normalizations. We then train a fresh RFT seed and expose all of its 100-step checkpoints only to this replacement set; no earlier RFT test result is eligible for the reported locked comparison.

H Canonical Training and Evaluation Protocol

The preliminary custom loop had one on-policy inner step, so its importance ratio was one and clipping inactive; it also used no explicit KL, mean token log probability, and top-four absolute-advantage response filtering. These are material departures. Table 10 removes all four in the primary study. Training logs confirm nonzero clipped regions and archive the KL, entropy, gradient norm, response length, reward standard deviation, and completion truncation rate at every logged step.

For each saved problem we keep 256 responses in original generation order. pass@1 is the exact mean correctness over all stored responses, which equals the expected first-position metric under exchangeability. pass@256 is the exact presence indicator. For intermediate K, 1,000 without-replacement response subsets are drawn using fixed seed 2027. Plurality ties use first occurrence. The primary paired analysis resamples problems, preserving all response-derived statistics within a problem. Training-run inference uses a hierarchical bootstrap that first samples seeds and then problems within seed. We do not treat 128,000 responses as 128,000 independent benchmark observations.

H.1 Paired generation-length sensitivity

We select 100 MATH-500 problems by the smallest hashes of a fixed prefix and problem text, without consulting response outcomes. Each of Base and the three canonical GRPO policies generates 256 responses per problem to 3,072 tokens within its native 4,096-token context. The paired short endpoint is the first 2,048 actual generated token IDs, not a retokenized text prefix. Both endpoints use the same offline decoder; original streaming text is retained separately. This is a paired sensitivity experiment, not an exact reproduction of the historical streaming renderer. Temperature 0.7, top-p 0.95, question-specific seeds, and the question set are fixed across policies. No answers or problems are discarded.

The structured-semantic residual remains 21 points for Base and 19–21 for GRPO (Table 11). We retain all three mappings and four accuracy endpoints for every policy in a 48-contrast family, using 20,000 paired problem-bootstrap draws and simultaneous max-T intervals. For the structured-

Table 11: Paired length sensitivity on the fixed 100-problem subset under structured-semantic mapping. Accuracy and residual columns are percentages or percentage points; cap is the percentage of responses length-terminated at 3,072 tokens. Each paired cell lists 2,048 / 3,072.

Policy	Oracle@256	Plurality@256	Residual	Cap (%)
Base	93 / 93	72 / 72	21 / 21	36.36
GRPO-42	92 / 93	72 / 73	20 / 20	30.37
GRPO-1	93 / 93	72 / 72	21 / 21	31.43
GRPO-2	93 / 92	73 / 73	20 / 19	30.62

Table 12: Canonical MATH-500 accuracy curves (%) for Base and GRPO seed 42. The main paper reports the three-seed aggregation.

Model	K=1	2	4	8	16	32	64	128	256
Base oracle	27.64	41.76	56.48	68.99	77.92	83.74	87.53	90.11	92.00
Base plurality	27.64	27.60	37.53	49.63	57.90	62.32	64.32	65.06	64.80
GRPO-42 oracle	31.95	46.97	61.29	72.30	79.66	84.45	87.63	89.94	92.00
GRPO-42 plurality	31.95	31.99	43.13	55.12	62.18	65.54	67.30	68.19	68.20

semantic residual change, the GRPO-42 and GRPO-2 intervals are $[-4.27, 4.27]$ and $[-4.00, 2.00]$ points; Base and GRPO-1 have no observed problem-level residual changes. A degenerate bootstrap interval in these unchanged cells does not establish population equivalence. Structured pass@1 decreases by 0.23–0.38 points across the four policies; the corresponding simultaneous intervals exclude zero, so extending a completion cannot be assumed to preserve its mapped correctness. Mean token use increases from 1,205 to 1,589 for Base and from 1,107–1,132 to 1,430–1,466 for GRPO. The still-high terminal cap rates limit the conclusion to this length extension: the experiment neither establishes uncapped performance nor length invariance. Full paired records and all contrasts are saved in `corrected_math_length_control.json`.

I Corrected Canonical Results

Seed 42 raises pass@1 by 4.3047 points and plurality@256 by 3.4 points. The absolute gap falls by 3.4 points because pass@256 is 92.0 in both rows. Table 14 reports problem-paired uncertainty and randomization tests. Only the normalized-gain difference is nonsignificant, and it points slightly upward rather than supporting a tax.

The high-K residual is not a contradiction to improvement. GRPO's gap is 20.34 points at $K = 64$, 21.75 at 128, and 23.8 at 256 because oracle coverage continues to rise after plurality has nearly saturated. The selection problem is therefore most visible when the entire curve is shown rather than only a single pass@1 or oracle endpoint.

J Clean 7B LoRA Scale Boundary

We separately repeat the corrected GRPO study with Qwen2.5-Math-7B; its exact revision is recorded in the artifact manifest. This boundary uses the same audited 10,000-problem training file, seeds 42/1/2, group size 16, two policy iterations, active clipping, KL coefficient 0.04, standard GRPO loss, learning rate 10^{-4}, rollout temperature 1.0, a 1,024-token training cap, and 400 optimizer steps. The deliberate difference is a rank-32 LoRA update (Hu et al., 2022) (alpha 64, dropout 0.05) over `q_proj`, `v_proj`, and `o_proj`; it is therefore not pooled with the full-parameter 1.5B primary comparison. Evaluation retains the same Qwen prompt, temperature 0.7, 2,048-token cap, MATH-500 problems, and 256 responses per problem. The Base row is a previously generated complete pool from the same immutable model revision and decoding protocol; it does not depend on training labels.

The conservative endpoint changes the apparent training contrast. Relative to Base, mean GRPO pass@1 rises by 23.85 points (hierarchical 95% CI $[20.69, 26.94]$), while plurality@256 changes by only $+0.07$ ($[-1.40, 1.53]$) and gap@256 by -0.87 ($[-3.00, 1.27]$). Pass@256 changes by -0.80

Table 13: Canonical GRPO seed-level MATH-500 results (%). SD is sample standard deviation across independent training runs.

Seed	pass@1	pass@256	Plur.256	Gap	G_{256}
42	31.95	92.0	68.2	23.8	53.27
1	31.27	91.2	69.0	22.2	54.90
2	31.97	91.8	68.0	23.8	52.96
Mean	31.73	91.67	68.40	23.27	53.71
SD	0.40	0.42	0.53	0.92	1.04

Table 14: Problem-paired comparison for Base versus canonical GRPO seed 42. Confidence intervals use 20,000 paired bootstraps; p uses 20,000 paired sign randomizations.

GRPO − Base	Estimate	95% CI	p
pass@1	$+4.30$	$[3.79, 4.84]$	< 0.00005
pass@256	0.00	$[-1.4, 1.4]$	1.000
plurality@256	$+3.40$	$[1.4, 5.4]$	0.00145
gap@256	-3.40	$[-5.8, -1.0]$	0.00990
G_{256}	$+1.92$	$[-0.87, 4.80]$	0.192

points ($[-2.27, 0.67]$). Complete-box mapping leaves 14.0–16.2-point gaps and the same absence of a stable gap direction. The released fallback instead yields the much larger Base gap and apparent reduction reported by the compatibility audit; this discrepancy is why it is not our cross-scale headline. Because model size and adapter geometry change together, the experiment tests transfer of the residual, not a parameter-count causal effect.

The disjoint 128/128-pool audit also transfers. Predicted/observed gap is 27.63/28.46% for 7B Base and averages 12.60/13.35% across the three GRPO seeds (mean prediction error -0.74 points). Full-pool strict recoverability rises from 62.6% to 76.73%, mean normalized entropy falls from 0.421 to 0.206, and modal margin grows from 20.48 to 48.18 points. Thus the categorical model remains calibrated while the trained policy makes correct answers substantially more modal.

The evaluator also exercised its fail-closed streaming invariant: one shard ended at 495/500 aggregate problems under its original wall time. Metrics were blocked; the five absent problems were generated from the same locked adapter and decoding configuration, and the final merge then verified 500 unique problems, exactly 256 responses each, and canonical reference order.

K Cross-Family Mistral-7B Boundary

To separate architecture transfer from Qwen-only scale transfer, we repeat the 7B LoRA boundary with Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), pinned to revision `c170c708c41dac9275d15a8fff4eca08d52bab71`. Training uses the same clean 10,000-problem file, seeds 42/1/2, group size 16, two policy iterations, GRPO loss, KL coefficient 0.04, rank-32 LoRA targets, 1,024-token cap, and 400 steps as the Qwen-7B boundary. A training-only 20-step stability screen, performed without benchmark generation, selects learning rate 10^{-5}: the transferred 10^{-4} rate reaches KL 733.8 by step 10, whereas 10^{-5} remains below 0.0013 through the screen. The frozen protocol records both outcomes and requires all final seeds regardless of direction.

Mistral's released tokenizer requires special care. The model receives one user message containing the problem and shared boxed-answer instruction in both training and evaluation. Evaluation calls `apply_chat_template(tokenize=True)` and passes those token IDs directly to vLLM. A human-readable `tokenize=False` representation does not round-trip to the same IDs; an incomplete engineering run using that path was stopped before metric computation and is isolated from every merged pool. The accepted merge verifies 500 unique reference-ordered problems and 256 responses per problem for Base and all three adapters.

Table 15: Clean-data Qwen-7B LoRA boundary under structured-semantic answer mapping on MATH-500 (%). SD is across three independently trained adapters.

System	p@1	p@256	Plur.	Gap	G_{256}
7B Base	40.96	92.6	77.8	14.8	62.40
GRPO-42	65.52	91.6	78.0	13.6	36.19
GRPO-1	66.81	91.8	78.4	13.4	34.92
GRPO-2	62.09	92.0	77.2	14.8	39.86
Mean	64.81	91.80	77.87	13.93	36.99
SD	2.44	0.20	0.61	0.76	2.56

Table 16: Cross-family Mistral-7B LoRA boundary under structured-semantic answer mapping on MATH-500 (%). SD is across three independently trained adapters.

System	p@1	p@256	Plur.	Gap	G_{256}
Base	11.43	76.0	27.2	48.8	17.81
GRPO-42	12.07	76.2	26.4	49.8	16.30
GRPO-1	11.83	77.0	25.8	51.2	15.84
GRPO-2	11.95	75.8	25.2	50.6	15.05
Mean	11.95	76.33	25.80	50.53	15.73
SD	0.12	0.61	0.60	0.70	0.63

Relative to Base, mean GRPO pass@1 changes by $+0.52$ points (hierarchical 95% CI $[0.22, 0.86]$), pass@256 by $+0.33$ ($[-2.20, 2.87]$), plurality@256 by -1.40 ($[-3.60, 0.73]$), gap by $+1.73$ ($[-1.60, 5.00]$), and normalized gain by -2.08 ($[-4.48, 0.27]$). Complete-box voting gives 47.6–49.4-point gaps. The residual is therefore not an artifact of the released fallback mapper, while every high-K training contrast spans zero. This establishes transfer of the measurement phenomenon, not a common GRPO direction or a controlled architecture effect.

A deterministic 4,096-response reservoir per policy finds 1.17% Base responses within eight tokens of the 2,048-token cap, versus 1.25%, 1.46%, and 0.56% for the three GRPO seeds. Median lengths are 424 versus 425/423/396 tokens. The large residual is therefore not produced by a common asymmetric truncation boundary.

The 128/128 split is also a useful boundary on the categorical estimator. The finite-sample plug-in underpredicts observed gap by 3.58 points for Base and 3.44–3.82 points across GRPO seeds; problem-level MAE is 20.8–21.8 points. Mistral produces 123–133 unique normalized answers per problem and effective support 80–90, making a 128-response histogram substantially sparser than in the Qwen study. The simpler strict-modal count is within 0.15–1.00 point, but it does not predict finite-K multinomial selection probability. We therefore retain the plug-in miss rather than extending the main paper's within-one-point Qwen claim to this family.

L Disjoint-Pool Predictive Validation

The following MATH-500 concentration and 128/128 prediction results use the released fallback mapper, not the conservative mappers of Table 1. Table 17 compares Base with GRPO-42.

For each of 100 seeded partitions, Pool A and Pool B contain 128 disjoint saved responses per problem. Pool A estimates the full normalized-answer histogram. We compute plug-in expected pass@128 and plurality@128 under fresh multinomial sampling from that histogram. Pool B then supplies the realized gap. No calibration parameter is fitted.

Prediction-minus-observation means are -0.81 points for Base and -0.38 for GRPO. The 2.5–97.5 percentiles across reused random partitions are $[-3.04, 1.59]$ and $[-2.39, 1.85]$ points. These are split-sensitivity ranges, not confidence intervals from independent models. Problem-level plug-in MAE is 12.06 points for Base and 11.11 for GRPO, so the result validates aggregate calibration, not exact labels for individual problems.

Table 17: Direct categorical statistics over all problems and disjoint 128/128 aggregate-gap validation. Gap columns are percentage points; other masses are probabilities.

Model	H_{norm}	e^H	Unique	Correct mass	Top-wrong	Margin	Pred. gap	Obs. gap
Base	0.607	47.52	110.64	0.276	0.124	0.153	24.23	25.04
GRPO-42	0.561	38.63	99.98	0.319	0.128	0.191	21.27	21.65

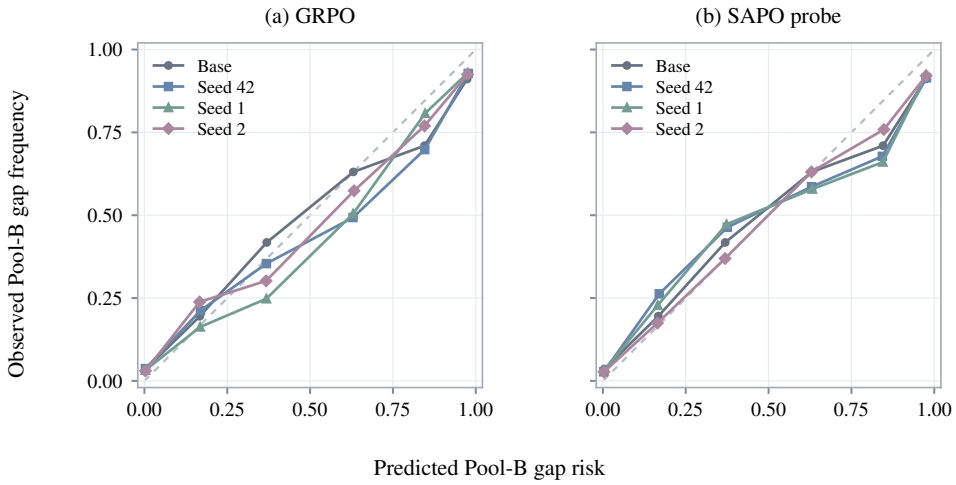

Figure 5: Pool-A predicted gap risk versus Pool-B observed gap frequency, separated by training family on shared scales. Base is repeated as a reference; colored curves show the three training seeds, not confidence bands. The dashed diagonal denotes perfect calibration. Fixed bins pool problem–partition pairs; split-level aggregate uncertainty is reported in text.

On the Base-defined hard solvable subset (366 problems), correct mass rises from 20.19% to 25.11%, top-wrong mass changes from 12.68% to 12.94%, and modal margin rises from 7.51 to 12.17 points. The strictly recoverable share rises from 62.30% to 67.21%; solvable-but-non-recoverable share falls from 36.89% to 30.33%. These measurements show beneficial concentration toward correct answers alongside the residual, and replace the old statement that wrong modes systematically become stronger.

Near-tie cases are explicitly uncertain. Pool-A-defined solvable problems with $|\widehat{\Delta}| \leq 0.05$ vary in count across partitions, and their conditional gap estimates have far wider ranges than the aggregate. Threshold sensitivity for $\delta \in \{0, 0.02, 0.05, 0.10\}$ and every calibration bin is included in the released JSON rather than summarized with a single optimistic error number.

L.1 Fixed-target sample complexity

To vary estimator data without moving the target, each of 20 seeded partitions reserves one fixed 64-response Pool B and uses nested disjoint Pool-A prefixes of 16, 32, 64, and 128 responses. Pool B and its plurality budget are identical across all four estimates within a partition. Table 18 reports means over reused partitions, not confidence intervals from independent model runs.

The Qwen aggregate happens to average out well even with 16 estimation responses, but its instance-level errors fall steadily with more data. Within Qwen-1.5B, the highest full-pool effective-support quartile has MAE 34.8/31.9 points at 16 responses and still 25.9/22.2 at 128 for Base/GRPO-42. Mistral's much broader support prevents such cancellation: signed aggregate underprediction contracts from $-13.2/-13.8$ to $-1.7/-2.2$ points. Thus the earlier Mistral miss is a predicted consequence of sparse histogram estimation, not evidence for a calibrated per-instance mechanism.

Table 18: Categorical prediction sample complexity (percentage points). Abs. err. is absolute aggregate prediction error; problem MAE is shown as Pool-A size $16 \to 128$.

Policy	Abs. err. 16	Abs. err. 128	Problem MAE
Qwen-1.5B Base	1.10	0.89	$19.2 \to 12.4$
Qwen-1.5B GRPO-42	1.07	0.95	$16.8 \to 10.8$
Qwen-7B Base	1.53	0.88	$14.9 \to 9.0$
Qwen-7B GRPO-42	2.24	0.90	$8.8 \to 6.4$
Mistral-7B Base	13.17	2.07	$30.4 \to 23.5$
Mistral-7B GRPO-42	13.78	2.40	$30.6 \to 23.2$

Table 19: Same-budget prediction at $|A| = 128$, $|B| = 64$. Each entry is plug-in / direct random-subset estimation. All losses are multiplied by 100; lower is better. Means over reused partitions are descriptive, not confidence intervals.

Policy	Aggregate abs. error	Problem MAE	Brier loss
Qwen-1.5B Base	0.96 / 0.90	11.94 / 10.68	6.61 / 6.77
Qwen-1.5B GRPO-42	0.90 / 0.95	10.76 / 9.73	6.23 / 6.43
Qwen-7B Base	0.77 / 0.83	8.94 / 7.93	5.09 / 5.20
Qwen-7B GRPO-42	0.62 / 0.57	6.38 / 6.06	3.92 / 4.11
Mistral-7B Base	1.80 / 1.46	23.85 / 23.14	14.68 / 15.62
Mistral-7B GRPO-42	1.56 / 1.25	23.33 / 22.22	14.14 / 14.73

Same-budget direct-estimation baselines. An additional fixed comparison uses the same labeled Pool A for three estimators: the categorical plug-in, the mean realized gap of disjoint 64-response blocks, and the mean gap of 500 random 64-response subsets drawn without replacement. The last estimator directly averages the finite-bank endpoint, unlike the plug-in's sampling with replacement. Every method predicts the same disjoint 64-response Pool B. We use 20 fixed partitions and Pool-A sizes 64 and 128; smaller pools cannot directly supply a 64-response block. Conditional modal ties in simulated counts are averaged uniformly, matching first occurrence in a random ordering. These are retrospective comparisons on existing banks, not independent replications.

The plug-in has no uniform error advantage (Table 19): direct estimation has lower problem MAE, while the plug-in's smoother risks have lower Brier loss. Aggregate error is similar on Qwen and favors direct subset estimation on Mistral. At $|A| = 64$, Mistral's plug-in aggregate error is 3.48/3.38 points versus 2.08/1.75 for direct subset estimation. Disjoint-block estimates show the same qualitative MAE/Brier trade-off; all endpoints and paired partition differences are released. Thus the categorical model supplies an interpretable prediction across budgets, but these data do not establish superior same-budget estimation or reliable instance-level routing.

An exact full-pool transition audit further separates oracle coverage from selection. Base has 40 unsolved, 136 missed, and 324 selected problems at $K = 256$; GRPO-42 has 40, 119, and 341. Of the 136 Base misses, 22 become selected, 108 remain missed, and 6 become unsolved. Conversely, 5 Base-selected problems become missed, and 6 Base-unsolved problems acquire a correct sample without a correct plurality. These paired counts are stored in `corrected_selection_state_audit.json`; they show that unchanged mean pass@256 can mask substantial problem-level movement in both directions.

M Complete Cross-System Ranking Audit

The main text reports the eight-system audit under the two conservative answer mappers. Table 20 summarizes the low-budget conservative and higher-budget fallback crossings. All budgets, including other differing point leaders with unresolved pairwise contrasts, appear in the released records. These descriptive crossings depend on the deployment contract and do not pass the full-audit simultaneous correction.

At $K = 8, 16$, RFT leads both oracle and plurality under both conservative mappers. At $K = 32$–256 their point leaders can differ, but no pairwise reversal

Table 20: Budgets with different point-estimate leaders. Regret is the best plurality accuracy minus that of the oracle leader (points); Res. is the number of 28 pairwise comparisons whose unadjusted paired 95% intervals exclude zero in opposite directions. The simultaneous reversal count is zero in every row; see Appendix F.

Mapper	K	Oracle / plurality leader	Regret	Res.
Released fallback	32	SAPO-42 / RFT	3.22	2
	64	SAPO-42 / RFT	2.46	4
	128	GRPO-2 / RFT	2.45	3
	256	Base, GRPO-42 / RFT	2.40	0
Complete-box	2	RFT / GRPO-42	4.48	6
	4	RFT / GRPO-42	2.45	6
Structured-semantic	2	RFT / GRPO-42	4.53	6
	4	RFT / GRPO-42	2.38	6

Table 21: Development-only checkpoint selection for fresh RFT seed 2027. Target GRPO pass@1 is 23.31%; step 100 is locked before test generation.

RFT step	Dev pass@1	$\lvert\text{RFT} - \text{GRPO}\rvert$
100	22.29	1.03
200	21.94	1.38
300	17.30	6.01
400	15.80	7.51
500	15.58	7.74
600	16.06	7.25
700	15.14	8.17
800	16.06	7.25

passes pointwise intervals. Released-fallback crossings pass pointwise intervals at $K = 32, 64, 128$ but not 256; none survives simultaneous correction. All 756 pair–mapper–budget records are retained in `corrected_system_ranking_audit.json`. The earlier `corrected_system_decision_audit.json` uses a Monte Carlo $K = 4$ estimate; exact low-budget endpoints and the simultaneous correction in `corrected_ranking_multiplicity_audit.json` supersede those estimates.

N MATCHED-PASS@1 REJECTION-SFT REFERENCE

The reference checkpoint is selected without MATH-500 access. Base and GRPO development pass@1 are 20.40% and 23.31%. Table 21 lists every checkpoint from the fresh RFT seed. The closest is step 100 at 22.29%, 1.03 points from GRPO and within the predeclared 1.5-point tolerance. Its problem-paired pass@1 difference is -1.03 points (95% CI $[-2.50, 0.45]$, randomization p=0.175). The job controller then, and only then, submits the locked MATH-500 generation.

Under the released compatibility endpoint, the locked test yields Table 22. RFT reduces the absolute gap by 4.0 points relative to GRPO-42 and raises normalized gain by 3.49 points; both paired intervals exclude zero. Pass@1 is nearly identical, while oracle and plurality contrasts are individually inconclusive. The complete $K \in \{1, 2, 4, 8, 16, 32, 64, 128, 256\}$ curve and all 20,000-trial paired results are stored in the accompanying JSON. Under complete-box-only voting, however, RFT and GRPO-42 have gaps of 19.4 and 20.8 points, so the contrast shrinks to -1.4 points. Structured semantic voting yields 19.6 and 19.8, a -0.2-point contrast; Table 6 reports the full sensitivity.

Development pass@1 is estimated from 16 responses per problem, so the matching procedure has sampling error. RFT and GRPO also use different training objectives and numbers of accepted trajectories. They are matched on held-out pass@1, not on compute.

The fixed-pool categorical audit provides a complementary distributional contrast. For GRPO-42/RFT respectively, normalized entropy is 0.561/0.582, effective support is 38.63/35.53, correct mass is 0.3195/0.3201, top-wrong mass is 0.1280/0.1106, modal margin is 0.1915/0.2094, and strict recoverability is 0.680/0.702. The fact that normalized entropy and effective support move in dif-

Table 22: Locked matched-pass@1 reference result on MATH-500 (%). The checkpoint and matching tolerance were fixed using the development set only.

System or contrast	pass@1	pass@256	Plur.@256	Gap@256	G_{256}
GRPO-42	31.95	92.0	68.2	23.8	53.27
Matched RFT	32.01	90.4	70.6	19.8	56.76
RFT − GRPO	0.06	−1.6	2.4	−4.0	3.49
Paired 95% CI	$[-1.19, 1.34]$	$[-3.2, 0]$	$[0, 4.8]$	$[-6.8, -1.2]$	$[0.16, 6.83]$
Randomization p	0.928	0.0940	0.0616	0.00860	0.0403

Table 23: Complete locked matched-pass@1 reference scaling curves on MATH-500 (%). Intermediate-K plurality uses 1,000 deterministic response subsets; endpoints use the complete response pool exactly.

K	1	2	4	8	16	32	64	128	256
GRPO pass@K	31.95	46.97	61.29	72.30	79.66	84.45	87.64	89.94	92.00
RFT pass@K	32.01	48.68	64.03	74.12	79.94	83.71	86.41	88.51	90.40
GRPO plurality	31.95	31.99	43.13	55.12	62.18	65.54	67.30	68.19	68.20
RFT plurality	32.01	32.02	45.59	59.88	66.24	68.48	69.34	69.77	70.60

ferent directions is not contradictory: the former averages $H/\log K$ with the same $K = 256$ for every problem, whereas the latter averages $\exp(H)$. Exponentiating before averaging weights high-entropy problems more strongly; these two averages need not rank systems alike. Neither is used alone as a success statistic.

Disjoint 128/128 validation predicts gaps 0.2121/0.1785 and observes 0.2172/0.1877 for GRPO-42/RFT, giving aggregate errors $-0.0051/-0.0092$. Problem-level MAEs are 0.1126/0.0863, and 100-split error ranges are $[-0.0281, 0.0214]$ and $[-0.0272, 0.0105]$. Thus the same pre-specified categorical model explains the direction of the RFT contrast out of pool while retaining substantial per-problem uncertainty.

O Released ProRL Boundary Case

We evaluate the released Qwen2.5-Math-1.5B base and ProRL-v2 checkpoints (Liu et al., 2025a) with temperature 0.6, top-p 0.95, maximum 8,192 new tokens, and $K = 64$. These settings follow the long-response release and differ from the canonical primary protocol, so the result is a boundary rather than a pooled baseline.

Paired ProRL-minus-Base differences are $+18.94$ points for pass@1, $+10.4$ for pass@64, and $+12.0$ for plurality@64, all statistically significant. Gap changes by -1.6 points with CI $[-3.6, 0.4]$ and $p = 0.181$. G_{64} changes by -11.38 points with CI $[-18.20, -4.28]$ and $p = 0.00105$. This is the clearest warning against interpreting normalized gain alone as harm: all three accuracy metrics improve while headroom-normalized gain falls.

Direct distributions also reject entropy-only reasoning. Correct-answer mass rises from 0.645 to 0.835 as normalized entropy falls from 0.210 to 0.051. Effective support falls from 7.30 to 1.76, strictly recoverable fraction rises from 0.738 to 0.860, and solvable-but-non-recoverable fraction falls from 0.054 to 0.040. Independent 32/32 prediction errors are -0.22 and -0.11 points. Wilson intervals for near-cap rates are $[15.10, 18.33]$ and $[0.63, 1.50]$ percent, so the trained model's gain is not an artifact of greater truncation.

P Frozen Practical-Selector Spectrum

The highest-performing tested selector is frozen before generation. A pinned Qwen2.5-Math-7B-Instruct checkpoint (Yang et al., 2024) greedily solves each MATH-500 question once with a 2,700-new-token budget. Under each answer mapper, its parsed answer selects the first matching answer cluster in the existing first-64 response pool; if absent, the same mapper's plurality winner is used.

Table 24: Released ProRL boundary case on MATH-500. Near-cap means response length at least 95% of the 8,192-token limit.

Model	pass@1	pass@64	Plur.@64	Gap	G_{64}	H_{norm}	e^H	Near-cap
Released Base	64.53	79.6	74.0	5.6	26.71	0.210	7.30	16.65%
ProRL-v2	83.47	90.0	86.0	4.0	15.33	0.051	1.76	0.98%

Table 25: Frozen independent-solver selector on MATH-500 at $K = 64$ (%). Match is the fraction of problems whose solver answer occurs in the sampled pool; residual is oracle minus selected accuracy.

Policy	Mapper	Match	Plur.	Gen-7B	Residual
Base	Fallback	89.8	62.6	77.6	10.0
Base	Complete box	89.8	68.2	77.8	9.8
Base	Structured	89.8	69.2	78.4	9.6
GRPO-42	Fallback	90.8	67.2	78.2	10.2
GRPO-42	Complete box	90.8	71.2	78.4	10.0
GRPO-42	Structured	90.8	72.0	79.0	9.8

Test answers are unavailable to solution generation, matching, and fallback. Model commit, source hashes, pool hashes, mandatory endpoints, and both policies are recorded before any output is inspected. Sixteen of 500 solver outputs reach the token cap.

For the registered compatibility endpoint, Gen-7B improves over plurality by 15.0 points for Base (paired problem-bootstrap 95% CI $[11.6, 18.4]$) and 11.0 for GRPO-42 ($[7.8, 14.2]$). Every mapper-specific improvement interval excludes zero. Its standalone mapped-answer accuracy is 77.0–77.8%. Under structured mapping, adding the pool changes accuracy by only 0.6 points for Base and 1.2 for GRPO-42. A post-hoc paired bootstrap over 20,000 draws, with simultaneous max-T intervals across all six policy–mapper increments, gives $[-1.40, 2.60]$ and $[-0.54, 2.94]$: neither resolves a benefit over the same solver alone. Base rescues 11 problems but loses 8; GRPO rescues 10 and loses 4. The large improvement over 1.5B plurality is therefore primarily the stronger solver's capability, not established complementary value from the candidate pool. A stronger selector nevertheless leaves a roughly ten-point gap. This supports a selector-quality spectrum, not an impossibility claim about future verifiers. Gen-7B is larger than the 1.5B policy and spends one additional long solve per problem; it is a strong-selector diagnostic, not a compute-matched efficiency comparison.

P.1 Selective solver routing: locked transfer boundary

We next ask whether the categorical margin can avoid invoking Gen-7B on every problem. This is post-hoc method development, not a preregistered primary experiment. On MATH-500 Base and GRPO-42, we sweep integer thresholds $\{-1, 0, 1, 2, 4, 8, 16, 32, 64\}$ for the difference between the two largest structured-semantic cluster counts. The router calls Gen-7B only when this margin is at most the threshold, uses its matching cluster with plurality fallback, and otherwise returns plurality. We choose the most accurate threshold whose mean call rate is at most 50%, breaking ties toward less compute. Before generating any OlympiadBench solver response, we freeze this rule, the model revision, the previously fixed 64-response pools, all four test policies, and mandatory reporting of every outcome.

Development selects threshold 8. Averaged over the two MATH policies, it raises accuracy from 70.6% for plurality to 76.9% while calling Gen-7B on 35.7% of problems; always invoking Gen-7B reaches 78.7%. Thresholds 16, 32, and 64 reach 78.2, 78.9, and 78.7%, respectively, but violate the development call budget. The full sweep and immutable input hashes are in the released JSON.

All four transfer point estimates exceed plurality, and the three-seed GRPO mean change is $+1.40 \pm 0.72$ points, but every paired interval includes zero. Moreover, the mean GRPO call rate is 50.73 ± 1.14%, versus 35.7% on development. The frozen test therefore provides weak directional evidence, not a successful routing claim: the raw count margin is not domain-calibrated. Solver outputs hit the length cap on 4.2% of problems, and 24.4–26.6% of invocations fall back because the solver

Table 26: Frozen cross-benchmark routing at $K{=}64$ (%). "Call" is the Gen-7B invocation rate; Δ is routed minus plurality with paired problem-bootstrap 95% CI; "Always" invokes Gen-7B for every problem.

Policy	Call	Plurality	Routed	Δ (95% CI) / Always
Base	50.8	39.2	40.8	$+1.6$ $[-0.6, 3.8]$ / 41.2
GRPO-1	52.0	40.0	42.0	$+2.0$ $[-0.2, 4.4]$ / 41.2
GRPO-2	50.4	41.2	41.8	$+0.6$ $[-1.6, 2.8]$ / 41.2
GRPO-42	49.8	41.0	42.6	$+1.6$ $[-0.6, 4.0]$ / 41.8

Table 27: Locked candidate-ranking judge on MATH-500 at $K{=}64$ under the structured-semantic mapper (%). Valid is the fraction of outputs containing a requested candidate ID.

Policy	Plurality	Judge	Change (95% CI)	Valid
Base	69.2	68.2	-1.0 $[-3.4, 1.4]$	51.4
GRPO-42	72.0	70.4	-1.6 $[-3.8, 0.6]$	52.4

answer does not match a sampled cluster. These two mechanisms are retained in the endpoint rather than filtered away. The result sharpens the practical boundary of the aggregate categorical account: it can predict population residuals accurately without yet supporting reliable instance-level compute allocation.

P.2 Held-out candidate-ranking judge

We separately test whether a generative judge can rank the same top-eight structured-semantic clusters without independently solving the problem. Each candidate receives a randomized letter ID and one representative response; the prompt contains no reference answer or policy identity. Three prompt variants are compared on 500 development questions disjoint from every locked benchmark. The development-selected variant includes cluster support counts and asks for one ID; invalid IDs fall back to plurality. The model revision, randomization, 512-token output cap, development choice, and locked MATH-500 and OlympiadBench endpoints are frozen before their outputs are opened (Stephan et al., 2024).

No improvement transfers to frozen OlympiadBench: changes for Base and GRPO seeds 1, 2, and 42 are -1.2, -1.6, -0.2, and -0.6 points, respectively. Long deliberation frequently exhausts the output cap before emitting an ID. A post hoc but fully disclosed one-token forced-choice follow-up makes every output parseable yet falls to 15.2/13.4% on MATH-500 and 7.2–10.8% on OlympiadBench. It is not presented as a second confirmatory run. Together the experiments show that this frozen recipe fails at candidate ranking, including after compliance is repaired; they do not bound stronger generative judges.

P.3 Held-out policy-aware verifier

We next adapt a selector to the evaluated rollout family without exposing any MATH-500 question or label during fitting, hyperparameter choice, or scoring. Before fitting, we froze a 500-question development set with 16 Base and 16 GRPO-42 responses per question. A two-normalization audit finds zero overlap with the 10,000 training prompts, MATH-500, and AIME-2024. A deterministic problem-key split assigns 400 questions to fitting and 100 to validation; all responses for a question stay on one side. The scorer is a pinned Qwen2.5-Math-7B backbone with a fresh scalar head and rank-16 LoRA (Hu et al., 2022). It learns eight within-question correct–incorrect pairwise rankings per eligible problem–policy pair for exactly three epochs. Separate Base-only and GRPO-only fits test cross-policy transfer; mixed-policy fitting is replicated with seeds 42, 1, and 2.

At locked test time, each answer mapper supplies the first representative of its eight largest clusters among the first 64 responses. The selected cluster maximizes verifier score plus $\lambda \log n_a$, where n_a is cluster count and $\lambda \in \{0, .25, .5, 1, 2, 4\}$ is chosen on validation questions before test scoring. Candidate construction and scoring omit ground-truth answers; the analysis joins labels only after all scalar scores have been written. The fixed 3,072-token input retains the first 512 and final 2,560

Table 28: Held-out adapted-verifier selection under structured-semantic mapping at $K = 64$ on MATH-500 (%). Mixed mean and SD are across three independent fits; all other fitted rows use seed 42.

Fitting rollouts / selector	Base pool	GRPO-42 pool
Base only	68.8	70.2
GRPO-42 only	64.8	64.2
Mixed, seed 42	69.2	69.4
Mixed, seed 1	70.2	72.4
Mixed, seed 2	68.0	69.8
Mixed mean	69.13 ± 1.10	70.53 ± 1.63
Plurality	69.2	72.0
Top-eight shortlist	85.2	84.6
Frozen Gen-7B	78.4	79.0
Oracle coverage	88.0	88.8

tokens, preserving the problem statement and final answer. We report the top-eight shortlist ceiling, paired problem-bootstrap intervals, all cross-policy transfers, and a hierarchical seed–problem interval for mixed training. A pre-fit smoke test found nonfinite outputs from a remote-code PRM under the available software stack, so the protocol was amended to the standard backbone plus a fresh head before any formal fitting or test score; every data split, endpoint, and training seed remained fixed.

Mixed-policy adaptation changes plurality accuracy by -0.07 points for Base (hierarchical 95% CI $[-2.53, 2.13]$) and -1.47 for GRPO-42 ($[-3.93, 0.67]$). Its oracle residuals are 18.87 and 18.27 points, while the shortlist-to-selected losses are 16.07 and 14.07. Base-only fitting is also indistinguishable from plurality on both pools; GRPO-only fitting is worse by 4.4 points on Base ($[-7.6, -1.4]$) and 7.8 on GRPO-42 ($[-10.8, -5.0]$). Thus policy-specific fitting does not solve ranking shift in this protocol, including in-distribution GRPO-to-GRPO evaluation.

The mapper audit again matters. Under complete-box voting, mixed adaptation changes Base/GRPO-42 by $-0.20/-1.73$ points, with both hierarchical intervals crossing zero. Under the released fallback it gains 5.33 points on Base ($[2.73, 8.00]$) but only 1.93 on GRPO-42 ($[-0.47, 4.20]$). We therefore do not promote the fallback-only gain to a general verifier result. Validation chose $\lambda = 2, 1$, and $2/4/2$ for Base-only, GRPO-only, and the three mixed fits under structured mapping. Training used 1,736/1,760 pairwise examples for the single-policy fits and 3,496 for each mixed fit; fewer than 0.09% of training sequences and 0.10% of scored test candidates were truncated. Every score pool contains exactly 500 unique problem hashes, records no answer field, and matches the frozen input and adapter manifests. The verifier's negative result is informative about this adaptation recipe, not an impossibility claim: Gen-7B demonstrates that a different selection paradigm performs much better.

A fail-closed pre-analysis audit caught one extra character in a copied rollout SHA-256 field. The immutable protocol, valid 64-character source hash, and an erratum recording their relationship are all released; the audit stopped before any accuracy output, and no split, data, model, or endpoint changed.

P.4 Official 72B outcome-reward selector

As a second frozen paradigm, Qwen2.5-Math-RM-72B scores the first response from each of the eight most frequent answer clusters under every mapper, using its released conversation format and scalar sequence reward. The union is scored once; the highest-reward mapper-specific candidate is selected. Cluster construction never uses correctness. This selector is excluded from our effectiveness comparisons pending interface validation. Under Transformers 4.44.2, its pinned model-card example yields 3.28125 instead of the documented 3.75, outside our prespecified absolute tolerance of 0.3, although all six fixed arithmetic pairs rank the correct response higher. The gate therefore stops before benchmark scoring. Earlier scores under Transformers 5.9.0 are withdrawn: reproducibility of those scores does not validate the interface or establish reward-model failure under policy shift.

Table 29: Validated PRM selectors on all 500 MATH-500 problems at $K = 64$ (%). Product is the prespecified primary aggregation.

			Terminal		Minimum		Product	
Policy	Mapper	Plur.	BoN	WSC	BoN	WSC	BoN	WSC
Base	Fallback	62.6	71.8	69.2	68.2	71.4	68.2	71.0
	Complete box	68.2	71.8	70.4	68.2	73.8	68.2	74.2
	Structured	69.2	72.6	71.8	69.6	75.0	69.6	75.4
GRPO-42	Fallback	67.2	71.0	71.6	69.4	72.6	69.4	72.6
	Complete box	71.2	71.0	72.2	69.4	73.8	69.4	74.6
	Structured	72.0	72.0	73.4	70.6	75.0	70.6	75.8

P.5 Public process-reward selector

Qwen2.5-Math-PRM-7B (Zhang et al., 2025) scores the unchanged first 64 responses for every MATH-500 problem, using its official chat template, double-newline step boundaries, and positive-class probabilities at step separators. Following the released paper, product aggregation is primary; minimum and terminal aggregation are mandatory sensitivities, not test-selected variants. Best-of-N chooses the highest response score. Weighted self-consistency (WSC) sums scores within answer categories; conservative mappers abstain on invalid outputs. Ties follow first occurrence. Log products and a common per-problem offset prevent numerical underflow without changing relative weights; an all-zero valid-weight vector falls back to plurality.

Remote-code compatibility is validated before benchmark scoring. The pinned official example is reproduced within 0.00432 absolute probability under Transformers 4.44.2, whereas the same weights fail this check under 5.9.0. Earlier scores from the incompatible stack are superseded, not interpreted as verifier failure. All six simple correct/incorrect arithmetic pairs are ranked correctly. On all 400 ProcessBench GSM8K examples (Zheng et al., 2024), the fixed first-step-below-0.5 rule gives 71.01% error localization, 95.85% all-correct recognition, and an 81.59% harmonic mean. These controls establish a working interface, not optimal selection on every distribution.

Product-WSC improves structured plurality by 6.2/3.8 points for Base/GRPO-42; the 95% simultaneous intervals are $[2.34, 10.06]/[0.48, 7.12]$ over all 36 policy–mapper–aggregation–selector contrasts (20,000 paired bootstrap draws). Gains also survive correction under both other mappers. Product-BoN alone does not significantly improve the structured endpoint. Thus verification combined with consensus recovers useful candidate-pool value, although 12.6/13.0 points remain below oracle. This is an existing public selector, not a new algorithm.

Scoring uses batches of one and a 4,096-token verifier-input cap. Four Base and three GRPO responses exceed this cap (0.0125/0.009375%); all are retained and scored on their prefixes. Three responses per policy retain no step separator and receive zero weight. On the 493 common problems with no capped verifier input, structured product-WSC gains remain 6.29/3.85 points. The artifact records raw/retained lengths, every step probability, all aggregations, positive controls, and the prescoring length amendment. These are capped-input selector results, not complete-trajectory verification.

P.6 Resource-accounted selector budgets

We replay the first $K = 8, 16, 32, 64$ responses of both corrected MATH-500 banks under structured-semantic mapping. The four pipelines are plurality, PRM product-weighted voting, the same standalone 7B solver, and pool-plus-solver matching. All $K = 64$ endpoints reproduce the earlier frozen results; smaller budgets are retrospective prefix diagnostics, not additional significance tests. Table 30 reports accuracy and separate resource dimensions.

Historical generator files retain decoded responses, but not the original token IDs or generation-token counts. We therefore retokenize their text with the generator tokenizer and the pinned solver tokenizer, recording file hashes and software versions. The resulting output-text proxy excludes discarded stop tokens and is not an exact reconstruction of generation work. Both prompt encodings average 98.94 tokens; the artifact reports unique prompt counts and K-fold counts separately rather than assuming prefix sharing. PRM input counts include its prompt and retained response prefix,

Table 30: Four-pipeline resource accounting on 500 MATH problems. Accuracies are percentages. Mean token counts are in thousands per problem. "1.5B out" is a re-encoded output-text proxy; PRM inputs are actual saved scorer-input counts, shown for all responses / eligible-only scoring. Standalone 7B has 77.8% accuracy and 0.656k output-text tokens, without a 1.5B pool.

Policy	K	Plurality	PRM vote	Pool+7B	1.5B out	PRM in
Base	8	54.8	60.2	66.8	9.46	10.33 / 5.20
	16	61.8	70.0	74.8	18.98	20.72 / 10.40
	32	67.0	72.8	77.2	37.99	41.47 / 20.98
	64	69.2	75.4	78.4	76.23	83.21 / 41.86
GRPO-42	8	57.8	63.6	68.4	9.00	9.88 / 5.57
	16	64.6	70.6	75.6	17.63	19.39 / 11.26
	32	72.6	74.6	79.0	35.25	38.76 / 22.22
	64	72.0	75.8	79.0	70.37	77.38 / 44.28

Table 31: Complete three-seed mean MATH-500 curves (%). No intermediate budget reverses the endpoint conclusion.

Mean curve	K=1	2	4	8	16	32	64	128	256
GRPO oracle	31.73	46.70	61.02	72.09	79.55	84.45	87.68	89.87	91.67
GRPO plurality	31.73	31.77	42.89	54.80	61.93	65.28	66.95	67.80	68.40
SAPO oracle	31.60	46.47	60.72	71.78	79.29	84.26	87.56	89.75	91.27
SAPO plurality	31.60	31.59	42.66	54.48	61.56	64.90	66.43	67.17	67.27

with the original 4,096-token cap. Eligible-only scoring removes calls that cannot contribute to weighted voting. Neither use of saved scores nor this filtering makes verification free.

Plurality incurs the generator resource vector; PRM voting adds verification; standalone 7B incurs only its own generation vector; pool matching incurs both generators. We never sum these heterogeneous token dimensions into a common FLOP or latency estimate. The standalone solver exceeds PRM voting in accuracy at all displayed budgets; this does not by itself establish a hardware-cost ordering. At $K = 64$, matching adds only 0.6/1.2 accuracy points to the solver, whose simultaneous intervals already include zero. Thus high oracle coverage does not establish that generating the candidate pool is worthwhile. Equal- hardware end-to-end timing and optimized serving would be required for a stronger efficiency claim. The released per-problem record reproduces all 4,000 policy–budget–question entries and keeps every budget, not only a favorable operating point.

Q SAPO DIAGNOSTIC PROBE

For each group, the implementation computes binary verifier rewards r_i and raw LOO credits c_i. Constant groups are assigned zero standardized reward. Otherwise each signal is group-standardized and mixed as

$$R_i^{\text{probe}} = (1 - \beta)z(r_i) + \beta z(c_i). \tag{20}$$

The schedule is fixed before evaluation: β increases linearly from 0 to 0.7 during the first 120 of 400 optimizer steps and remains 0.7 thereafter. The mixed signal enters the same canonical GRPOTrainer configuration as Table 10. Thus differences can diagnose whether selector-aware credit moves the learned distribution, but Proposition 1 does not imply monotonic improvement.

All three SAPO seeds use the same predetermined schedule and 400-step budget. We report the complete plurality curve, categorical statistics, AIME transfer, response lengths, and seed-level hierarchical intervals. No checkpoint or β is selected on MATH-500. In particular, the preliminary result in which an early tuned checkpoint appeared better and a later checkpoint reversed is not presented as confirmatory evidence.

Every interval spans zero, and the mean SAPO plurality curve is below GRPO for all $K \geq 2$. A deterministic reservoir audit of 2,048 responses per run finds near-2,048-token-cap fractions of 31.2/33.7/31.3% for GRPO and 31.2/30.6/31.4% for SAPO (seeds 42/1/2). The locked comparison is therefore not explained by a systematic truncation imbalance.

Table 32: Locked SAPO results and hierarchical 95% intervals for SAPO minus matched-seed GRPO (%).

Run	pass@1	pass@256	Plur.256	Gap	G_{256}
SAPO-42	31.83	91.4	67.6	23.8	52.47
SAPO-1	31.59	91.4	66.8	24.6	51.47
SAPO-2	31.37	91.0	67.4	23.6	52.50
Mean	31.60	91.27	67.27	24.00	52.14
SD	0.23	0.23	0.42	0.53	0.59
Δ vs. GRPO	-0.13	-0.40	-1.13	$+0.73$	-1.57
Hier. CI low	-0.62	-1.60	-2.80	-1.47	-4.08
Hier. CI high	0.35	0.80	0.47	2.93	0.89

Table 33: OlympiadBench-Text500 full-pool results at $K = 64$ (%). Box/Struct are the gap under complete-box and structured-semantic voting; Cap is the fraction ending at the 2,700-token generation limit.

Policy	pass@1	Oracle	Plurality	Gap	Box gap	Struct. gap	Cap
Base	20.01	65.0	38.8	26.2	25.4	25.8	19.93
GRPO-1	21.36	65.6	39.0	26.6	25.2	25.4	16.85
GRPO-2	21.88	65.0	41.0	24.0	23.8	23.8	16.20
GRPO-42	22.10	67.2	40.6	26.6	25.8	25.8	16.37
GRPO mean	21.78	65.93	40.20	25.73	24.93	25.00	16.47
GRPO SD	0.38	1.14	1.06	1.50	1.03	1.06	0.33

The disjoint-pool predictive audit also transfers to all six trained policies. Across seeds, GRPO's plug-in predicted and held-out observed gaps average 21.63% and 22.01%, versus 22.03% and 22.56% for SAPO; mean prediction errors are only -0.38 and -0.53 points. The underlying distributions barely move: normalized entropy is 0.5641 versus 0.5648, top-wrong mass 0.1275 versus 0.1277, and modal margin 0.1898 versus 0.1883. Strict recoverability instead falls from 68.13% to 67.00%. Thus the null endpoint result is consistent with the probe failing to materially reshape the categorical competition, rather than an evaluator failure on one policy.

R OLYMPIADBENCH-TEXT500 REPLICATION

We freeze this replication before model generation. From OlympiadBench's English split (He et al., 2024), 581 items meet all eligibility rules: open-ended competition mathematics, text only, one answer, an officially machine-gradable answer type, and a subfield in Algebra, Geometry, Combinatorics, or Number Theory. A seed-2027 sample fixes 500 question IDs. The manifest records the dataset fingerprint, selected IDs, four shard hashes, and official evaluator commit. Exact overlap with the 10,000 RLVR training questions is zero under both whitespace-removal and alphanumeric normalization.

All four policies generate 64 responses at temperature 0.7 and top-$p = 0.95$. An initial engineering pilot at 2,048 new tokens was discarded before any correctness result because 19–21% of responses hit that limit. Every policy was regenerated at 2,700 new tokens within the checkpoint's native 4,096-token context. The pinned official symbolic judge is applied identically to all policies with a two-second fail-closed call timeout; zero comparison calls time out. We report the remaining difference in length-stop rates and test a sensitivity that makes length-stopped responses abstain while retaining them as pass@1 errors.

For GRPO-42 minus Base, paired 95% intervals are $[1.54, 2.66]$ points for pass@1, $[-0.4, 4.8]$ for oracle@64, $[-0.2, 4.0]$ for plurality@64, and $[-2.8, 3.6]$ for gap@64. Thus individual-sample accuracy improves, the residual replicates, and its training-associated direction remains unresolved. When length-stopped responses abstain, Base and mean-GRPO gaps are 25.0 and 25.13 points. Mapper and truncation sensitivities therefore preserve the cross-benchmark residual.

Table 34: Frozen MBPP diagnostic at K=64 (%). Selected uses one visible unit test; oracle uses only held-out tests. GRPO mean and SD use the three canonical math-GRPO seeds.

Policy	pass@1	Oracle	Selected	Residual
Base	4.46	71.21	64.20	7.00
GRPO-42	4.80	71.21	65.37	5.84
GRPO-1	4.75	69.26	66.15	3.11
GRPO-2	4.86	73.15	67.32	5.84
GRPO mean	4.80	71.21	66.28	4.93
GRPO SD	0.06	1.95	0.98	1.57

The finite-sample account also transfers. In disjoint 32/32 splits, plug-in prediction versus observed gap is 20.69/21.96% for Base and 20.17/21.60, 19.69/20.76, and 21.14/23.15% for GRPO seeds 1, 2, and 42. Prediction errors range from -1.07 to -2.01 points; problem-level MAE is 17.7–19.1 points. This is aggregate predictive calibration, not reliable per-problem classification.

S Frozen MBPP Code-Output Diagnostic

To test whether the evaluation distinction depends on mathematical answer normalization, we freeze a non-mathematical diagnostic on the 257 sanitized MBPP test tasks (Austin et al., 2021). For each task, the prompt exposes the problem statement and only the first listed unit test. The remaining two or more tests are withheld from generation and selection. Each policy samples $K = 64$ responses at temperature 0.7 and top-$p = 0.95$ with a 768-token cap. A sandboxed evaluator parses Python code, rejects unsafe abstract-syntax-tree constructs, and enforces process resource limits. The practical selector picks the first candidate with the largest number of public tests passed; correctness and oracle coverage require all hidden tests to pass. The benchmark, protocol, evaluation code, and endpoints were hashed before generation.

Every policy has a positive oracle–selector residual: Base's paired problem-bootstrap interval is $[3.89, 10.51]$ points; the three GRPO intervals are $[3.11, 8.95]$, $[1.17, 5.45]$, and $[3.11, 8.95]$. The three-seed GRPO mean does not differ significantly from Base in oracle coverage (difference 0.00, CI $[-3.50, 3.50]$) or selected accuracy ($+2.08$, CI $[-1.82, 6.10]$). Only 13.4–14.6% of generations contain extractable code, explaining the low single-sample accuracy and making the high oracle at 64 draws a demanding selector setting rather than a code-capability result. This experiment therefore transfers the coverage-versus-selection distinction to executable outputs with held-out behavioral correctness. It does not test code-trained RLVR, and possible benchmark memorization is not ruled out.

S.1 Same-backbone 14B code-RL boundary

We repeat the frozen public-test/hidden-test contract with the pinned DeepCoder-14B-Preview code-RL checkpoint and its declared initializer, DeepSeek-R1-Distill-Qwen-14B (Luo et al., 2025). The model card describes distributed code RL and names TACO-Verified, PrimeIntellect SYNTHETIC-1, and LiveCodeBench as training sources; it does not establish the absence of MBPP from pre-training. Both checkpoints use the released reasoning chat template, $K = 32$, temperature 0.6, top-$p = 0.95$, and identical task order and evaluator.

Table 35: Frozen same-backbone 14B MBPP comparison (%). The 4K run is a token-cap truncation diagnostic; 16K tokens is the predeclared cap-triggered extension. Gap intervals are paired problem-bootstrap 95% CIs.

Cap	Checkpoint	p@1	Oracle	Selected	Gap (CI)	Capped
4K	Initializer	82.93	94.16	90.66	3.50 [1.56,5.84]	14.52
	Code RL	77.04	91.05	89.11	1.95 [0.39,3.89]	34.52
16K	Initializer	86.84	96.11	91.44	4.67 [2.33,7.39]	1.12
	Code RL	86.43	95.33	93.00	2.33 [0.78,4.28]	5.19

Table 36: Strictly filtered AIME 2024 I/II results (30 problems, %).

Run	pass@1	pass@256	Plur.256	Gap
Base	6.80	63.33	23.33	40.00
GRPO-42	7.30	56.67	20.00	36.67
GRPO-1	7.04	53.33	20.00	33.33
GRPO-2	6.86	53.33	20.00	33.33
GRPO mean	7.07	54.44	20.00	34.44
GRPO SD	0.22	1.92	0.00	1.92
SAPO-42	6.91	56.67	23.33	33.33
SAPO-1	7.01	63.33	20.00	43.33
SAPO-2	6.85	63.33	20.00	43.33
SAPO mean	6.92	61.11	21.11	40.00
SAPO SD	0.08	3.85	1.92	5.77

The 4K maximum cap rate exceeds the frozen 10% boundary-only threshold, so its apparent endpoint directions are not transfer evidence. Before examining any 16K outputs, we froze the extension and classification rule: below 5% capped is clean, 5% through 10% is qualified, and above 10% is boundary-only. The 16K maximum of 5.19% is therefore qualified. Code-RL minus initializer changes at 16K are -0.41 points for pass@1 (95% CI $[-1.57, 0.63]$), -0.78 for oracle ($[-2.72, 1.17]$), $+1.56$ for selected accuracy ($[-0.78, 3.89]$), and -2.33 for the residual ($[-5.06, 0.39]$). None resolves a training-associated direction, while each checkpoint's residual interval excludes zero.

The first extension launch produced no usable comparison because two L40S nodes reported uncorrectable ECC errors and some shards were incomplete. We inspected only scheduler states, exceptions, row counts, and throughput, froze a re-execution note, discarded all L40S files, and reran every model–shard cell on H100 with batch size one. All 16 shards then completed with exactly 257 unique tasks per checkpoint. Protocol, errata, model revisions, sample hashes, and the failed-node record are released. This comparison adds scale and code-specific RL transfer, but remains a public-checkpoint boundary rather than a controlled causal estimate of code RL.

T AIME 2024 STRESS TEST

AIME 2024 contains only 30 problems, so seed and problem uncertainty are large. The cached AI-MO validation split (AI-MO, 2024) concatenates 2022–2024 (90 problems); its URL field was therefore filtered for 2024 before metrics were recomputed. This audit changed the originally mislabeled 90-problem result and is enforced by a 30-row assertion in the loader.

GRPO does not improve the high-K accuracies in this stress test: both oracle coverage and plurality are lower. The smaller gap is therefore not a method success, underscoring why gap must be reported beside both endpoints. SAPO recovers some oracle coverage relative to GRPO but its plurality values are 23.33/20.00/20.00%, providing no stable selection improvement. We treat every AIME row as a low-powered cross-benchmark stress test and do not use it for model, checkpoint, or hyperparameter selection.

U COMPUTE, SOFTWARE, AND REPRODUCTION

Training and evaluation run on a managed Slurm cluster using NVIDIA A40, L40S, H100, and H200 GPUs as recorded per job. The 1.5B GRPO, SAPO, and RFT scripts request one H200, eight CPU cores, and 96 GB host memory; the 7B script requests one H100 or H200 as recorded, eight cores, and 96–128 GB. The final 14B MBPP generation uses one H100 per model–shard cell, eight modulo shards per checkpoint, and batch size one; evaluation uses 32 CPU workers. The operating-system image and a complete package lock were not archived. Canonical GRPO and SAPO training use the same requested resources and wall-clock limit. Evaluation is sharded only by problem; deterministic merge code checks exactly 500 unique problem identifiers, exactly 256 responses per problem, and reference order before computing metrics.

The artifact includes per-run metrics, ordered problem-local categorical answer IDs, and machine-readable generation protocols sufficient to reconstruct every reported selector curve and both mapper sensitivities. It also provides the balanced parser and its unit tests for nested braces and incomplete trailing expressions, together with data-quality, label-change, leakage, and response-length audits.

Training and evaluation shell scripts specify the Slurm resource requests. Analysis files cover paired comparisons, hierarchical bootstraps, disjoint-pool prediction, and answer-mapper sensitivity. Plotting scripts are accompanied by machine-readable figure inputs. The frozen candidate-judge, sample-complexity, decision-audit, selective-router, and MBPP/14B protocols and analyses are included.

The artifact includes a deterministic 50-problem raw-text audit subset for Base and GRPO-42, with all 256 responses and expected outputs for every mapper; the compact records regenerate byte-for-byte. The remaining raw generations and checkpoints are not duplicated because of archive size. Public model revisions, generation commands, compact labels, and all derived statistics are included; regenerating the full raw pools remains hardware- and software-environment dependent. The code artifact intentionally contains no paper source or review PDF.

The test suite currently checks balanced extraction, incomplete-answer handling, validity-aware voting and invalid-response denominators, the exact $K = 1$ metric identity, exact full-pool PRM evaluation, PRM $K = 1$ identity, top-cluster RM reconstruction, ground-truth-free generative-selector matching and fallback, official symbolic-category handling, and rejection of partial, duplicate, or wrong-K streaming pools. Additional tests check exact ordered-subset selection, invalid-abstention identities, and paired-bootstrap stratum weights. The public-bank analysis includes a pinned downloader and an independent scalar replay of every full-bank endpoint, alongside its complete simultaneous comparison family.

V High-Budget Decisions on Public Response Banks

Source and analysis status. We use the `harimo/scorio-lite` banks released by Hariri et al. (2026a), pinned to revision `8a00506b9f016b7136b366f105f137d5947d5093`. This is secondary analysis of independently generated public data. The source paper's aggregate findings were available when we froze our plan; the plan was fixed before loading response-level outcomes, not before the source publication. The source study already distinguishes discovery from selection. Our additional analysis tests cross-configuration ranking reversals under a joint comparison family and an invalid-vote counterfactual.

Fixed evaluation contract. All four configurations and both blocks are retained: 186 competition-math problems from five contests, and 3,600 SuperGPQA problems across 72 fields, each with seeds 0–79. The three gpt-oss effort levels share one checkpoint; they are configurations, not independent trained models. We align math by task and problem ID, and SuperGPQA by UUID. Stored problem prompts and references align exactly across configurations. Literal extracted categories are only stripped of surrounding whitespace; empty/null answers are invalid. Source correctness is used to grade choices, never to merge categories or choose candidates.

Plurality either abstains on invalid outputs or lets them vote as one wrong category. Both break ties by first occurrence. Reference-free BoN maximizes the mean of the source's three expected criterion scores, again breaking ties by first occurrence. Missing scores receive negative infinity; if every score is missing, the rule falls back to abstaining plurality. No scores are missing here. We exclude the source's reference-assisted CompassVerifier. All source-capped and aborted attempts remain in the denominator.

Sampling and simultaneous inference. At $K = 80$, evaluation is exact in seed order. For $K = 32, 64$, each problem uses 1,000 ordered subsets without replacement, shared across configurations; oracle coverage uses its exact hypergeometric expectation. The fixed seed is 20270924. We bootstrap paired problems within each math task or SuperGPQA field, preserving stratum sizes, with independent resampling across blocks. The complete family is six pairs times three budgets times four endpoints times two blocks, or 144 scalar contrasts. The 20,000 centered standardized bootstrap replicates give a max-T critical value of 3.28594. A resolved reversal requires opposite

Table 37: Complete external $K = 80$ endpoints (%). "Vote" treats invalid outputs as one wrong category. BoN uses only reference-free scores. Qwen denotes Qwen3.6-35B-A3B; effort rows denote gpt-oss-20b configurations.

Block	Configuration	Oracle	Plurality	Vote	BoN
SuperGPQA	Qwen	87.58	66.47	66.44	65.81
	Low effort	84.81	40.25	40.03	45.61
	Medium effort	85.08	46.00	45.94	50.11
	High effort	81.94	47.89	47.67	50.69
Math	Qwen	94.62	89.25	89.25	86.56
	Low effort	72.58	45.70	45.70	58.06
	Medium effort	91.94	80.11	80.11	75.81
	High effort	93.55	85.48	85.48	81.72

Table 38: SuperGPQA configuration contrasts at $K = 80$ (percentage points). Intervals control the full 144-contrast family approximately, not just these displayed contrasts. Positive values favor the first setting.

Contrast	Endpoint	Difference	Simultaneous 95% CI
Low − High	Oracle	$+2.86$	$[0.91, 4.81]$
	Plurality	-7.64	$[-9.97, -5.31]$
	Invalid-vote plurality	-7.64	$[-9.97, -5.31]$
	Reference-free BoN	-5.08	$[-7.51, -2.65]$
Medium − High	Oracle	$+3.14$	$[1.65, 4.63]$
	Plurality	-1.89	$[-3.39, -0.39]$
	Invalid-vote plurality	-1.72	$[-3.24, -0.20]$
	Reference-free BoN	-0.58	$[-2.70, 1.53]$

signs in both simultaneous intervals. Inference is approximate and conditional on these banks, grading, verifier scores, and fixed subset draws; it does not include new-generation variability.

Results and limits. Tables 37–38 give the complete full-bank endpoints and the gpt-oss contrasts against high effort. Low–high reversals resolve for all three selectors at all three budgets; medium–high reversals resolve for both plurality rules, while BoN remains unresolved. These share systems and response banks, so multiple budgets and rules are sensitivities, not independent replications. Among the three gpt-oss settings, oracle chooses medium effort at $K = 80$ whereas plurality chooses high effort, losing 1.89 points. Among all four systems, Qwen wins under every endpoint: we do not claim a mistaken overall winner or a cross-backbone reversal. The math block has no point reversal. These results support configuration-specific decision risk, not universal ranking instability or a causal RLVR training effect.

Invalid-answer rates on SuperGPQA are 0.073%, 1.917%, 1.403%, and 1.936% for Qwen and low/medium/high effort, respectively. The low–high plurality contrast is unchanged by the invalid-vote control; medium–high changes only 0.17 points. This rules out abstention as a necessary mechanism for these particular reversals. No invalid answer is marked correct, and an independent scalar replay matches every $K = 80$ problem-level endpoint exactly. No literal category has conflicting source correctness within a problem/configuration.

Compute and interfaces remain important qualifications. Mean completion tokens on SuperGPQA are 4,682 for Qwen and 355/1,767/10,671 for the gpt-oss effort levels; capped fractions are 0.077% and 0/0.00035/1.110%. Seven medium-effort attempts abort and remain included. On math, mean tokens are 34,382 and 1,781/12,341/37,264, with capped fractions 4.54% and 0/0.22/16.12%. These are fixed-sample comparisons, not token- or FLOP-matched recommendations. Source verifier inputs differ: gpt-oss uses its final channel, whereas Qwen uses full text. Public benchmark contamination and source judging errors cannot be excluded. The artifact releases all endpoints, contrasts, per-problem records, file hashes, the frozen plan, and replay programs, including null results; no pair, parser, budget, or family was changed after inspection.

Table 39: Candidate-set decisions on SuperGPQA at $K = 80$. Loss is the selected-accuracy difference between the selector's empirical best configuration and the oracle-chosen configuration, in percentage points. These are descriptive views of the same banks, not independent replications or equal-compute comparisons. All 11 non-singleton subsets are in the artifact.

Candidates	Oracle choice	Plur. choice	BoN choice	Plur. loss	BoN loss
Low, high	Low	High	High	7.64	5.08
Three efforts	Medium	High	High	1.89	0.58
All four	Qwen	Qwen	Qwen	0.00	0.00

W Question-Disjoint Select-Then-Test Pilot

We replay a complete system-selection workflow on the public banks rather than choosing a system from its test accuracy. This is a retrospective pilot: the banks and their full-data aggregate outcomes were already studied. The split, candidate sets, budgets, selectors, selection criteria, and complete contrast family are fixed before computing this pilot's split-specific results. It is not a preregistered independent confirmation.

Within each original field/task stratum, we sort question identifiers by a fixed SHA-256 key and allocate the first half (rounded down) to development. The remainder is test: 1,800/1,800 SuperGPQA and 92/94 math questions. No response to a test question enters development. Candidate sets are all four configurations and the three gpt-oss effort settings; neither excludes a system based on this pilot's results. Budgets are $K = 32, 64$, with both invalid voting and invalid abstention. Each selection criterion observes the same 80 responses per development question and system. This equalizes observed sample count, not token cost, FLOPs, or GPU time.

Four rules maximize development pass@1, oracle coverage, direct selector accuracy, or categorical predicted selector accuracy. Oracle uses the exact finite-bank hypergeometric expectation; direct evaluation uses the existing 1,000 ordered subsets without replacement. The categorical rule uses the literal empirical histogram with no smoothing, averaging 16,384 IID multinomial draws and exact conditional first-occurrence tie weights. It targets with-replacement population sampling, unlike the finite-bank direct estimator. Neither oracle budget exceeds the 80 observed responses. Ties between systems use a fixed candidate order. Development choices and scores are persisted before computing test contrasts; conflicting source grades within a literal category or invalid-yet-correct responses abort the replay instead of triggering correctness-based category merging.

The primary contrast is test plurality accuracy of the direct-selector choice minus that of the oracle choice, for three effort settings on SuperGPQA at $K = 64$ with invalid voting. Paired, field/task-stratified bootstrap inference uses 10,000 draws and a centered maximum-statistic correction across all $2 \times 2 \times 2 \times 2 \times 3 = 48$ contrasts: blocks, candidate sets, budgets, selectors, and non-oracle selection rules. The critical value is 2.3203. Intervals condition on the fixed development choices, response banks, source grades, and Monte Carlo draws. They do not cover repeated development selection or previously inspected-data reuse. Hindsight test-best regret is separately reported in the released JSON, not treated as a prespecified paired contrast.

For the primary comparison, test accuracy is 47.86% for high effort and 45.54% for medium, a 2.32-point gain. All three non-oracle rules agree, so this result does not establish categorical prediction's advantage over pass@1 or direct evaluation. Including Qwen leaves SuperGPQA choice unchanged. On math, the four-system $K = 64$ choice differs, but its test contrast is unresolved. Thus the pilot supports a question-disjoint consequence of effort-level selection while preserving both stable-choice and inconclusive scenarios. Its reused-data status requires new questions or generations before a prospective replication claim.

X Finite-Information Identification and Decision Certificates

Oracle curves specify total correct mass, but not its fragmentation into eligible categories or competition from wrong answers. We ask what selected accuracy is possible when only some of this information is retained. These are population identification questions. Substituting empirical frequencies

Table 40: Question-disjoint retrospective selection. "Other" denotes pass@1, direct selector evaluation, and categorical prediction, which choose the same system in every declared context. Differences are Other minus Oracle test accuracy, in points, with 48-contrast simultaneous intervals. Q, M, H denote Qwen, medium effort, and high effort. Grouped rows have identical choices and contrasts; all 48 scalar contrasts are released, including zeros.

Block	Candidates	K	Invalid	Oracle	Other	Test difference [CI]
SuperGPQA	All four	32,64	Both	Q	Q	0.00 [0.00,0.00]
	Three efforts	32	Votes	M	H	2.25 [1.01,3.49]
	Three efforts	32	Abstains	M	H	2.53 [1.31,3.75]
	Three efforts	64	Votes	M	H	2.32 [0.91,3.73]
	Three efforts	64	Abstains	M	H	2.60 [1.22,3.98]
Math	All four	32	Both	Q	Q	0.00 [0.00,0.00]
	All four	64	Both	H	Q	3.92 [−1.29,9.13]
	Three efforts	32,64	Both	H	H	0.00 [0.00,0.00]

answers the same questions for an empirical distribution; it does not yield confidence coverage for an unknown population.

Extremal wrong distributions. Let $J_K(c, q; \rho)$ be selected correctness for correct-category vector c, wrong vector q, and abstention mass ρ, with first-occurrence ties. Fix the correct counts, their largest count $r > 0$, the number t attaining r, and every wrong count except two. Let these two have combined count ℓ, with split $n, \ell - n$, and let u other wrong categories tie at r. If another wrong count exceeds r, the conditional win probability is zero. Otherwise it is

$$f(n) = \frac{t \mathbb{1}[n \le r,\ \ell - n \le r]}{t + u + \mathbb{1}[n = r] + \mathbb{1}[\ell - n = r]}.$$

This function is symmetric and nondecreasing toward $\ell/2$, hence a nonnegative linear combination of indicators of central intervals $[h, \ell - h]$. Conditional on ℓ, the split is $\mathrm{Bin}(\ell, v)$, where $v = q_a/(q_a + q_b)$. For $1 \le h \le \ell/2$,

$$\frac{d}{dv} \Pr(h \le N \le \ell - h) = \ell\{\Pr(\mathrm{Bin}(\ell - 1, v) = h - 1) - \Pr(\mathrm{Bin}(\ell - 1, v) = \ell - h)\}.$$

The two binomial coefficients coincide, so this derivative is nonnegative for $v \le 1/2$; symmetry handles $v \ge 1/2$. The conditioning variables have unchanged distributions when $q_a + q_b$ is fixed. Thus balancing wrong masses cannot reduce J_K, including when correct answers occupy multiple categories or invalid responses abstain. This is the relevant specialization of multinomial majorization (Rinott, 1973), not a new general majorization principle.

For total wrong mass W and maximum w, any sorted feasible vector has first-h sum at most $\min(hw, W)$. The packed vector attains these upper bounds and majorizes every feasible vector. With at most m wrong categories, retain one maximum w; balancing the other $m - 1$ coordinates gives the least concentrated feasible vector. This proves Proposition 2. If support is unrestricted, increase m to obtain the stated supremum. Zero remainders are omitted; $m = 1$ is deterministic given $W = w$; $W = 0$ gives $1 - \rho^K$. The finite-m case requires feasibility $W/m \le w \le W$. Countable wrong supports follow by truncation: moving a tail of mass ε changes a K-draw law by at most $K\varepsilon$ in total variation. Observed category count is never assumed to bound population support.

What correctness alone identifies. For one correct category of mass p, all answers voting, and $C \sim \mathrm{Bin}(K, p)$, the minimum is $\Pr(C > K/2) + \frac{1}{2}\mathbb{1}[K \text{ even}] \Pr(C = K/2)$. The supremum is

$$U_K(p) = 1 - (1 - p)^K - (K - 1)p(1 - p)^{K-1}.$$

The lower endpoint concentrates all wrong mass; the upper diffuses it so that $C \ge 2$ wins, while $C = 1$ wins with probability $1/K$. If correct-label fragmentation is also unknown, the lower endpoint instead becomes $p^K + (K-1)(1-p)p^{K-1}$: concentrate wrong mass and diffuse correct mass, equivalently complement the single-wrong-atom upper endpoint. Pooling correct categories cannot hurt correct selection, so the same $U_K(p)$ remains the supremum. Under unknown invalid abstention, the upper endpoint becomes oracle coverage, since all wrong outputs could abstain. Neither ordinary-voting endpoint is silently reused under a different validity rule.

Table 41: SuperGPQA development-distribution identification ranges at $K = 64$ with invalid outputs voting (%). These are not statistical confidence intervals. Full categorical prediction is shown for comparison.

System	Correct mass	Correct structure	Plus wrong maximum	Full
Qwen	$[40.13, 85.21]$	$[64.84, 85.21]$	$[66.78, 67.10]$	66.98
Low	$[7.45, 77.34]$	$[32.57, 77.34]$	$[39.55, 40.94]$	40.48
Medium	$[11.61, 78.66]$	$[41.24, 78.66]$	$[45.64, 46.49]$	46.13
High	$[17.05, 75.24]$	$[44.43, 75.24]$	$[47.36, 47.85]$	47.64

Exact finite-budget calculation. For a possible modal count $r \geq 1$ define

$$B_a(z, x) = \sum_{n=0}^{r-1} \frac{(az)^n}{n!} + x\frac{(az)^r}{r!}, \quad C_r = \prod_{a \in c} B_a, \quad Q_r = \prod_{a \in q} B_a.$$

Then

$$J_K(c, q; \rho) = K! \, [z^K] \sum_{r=1}^{K} \int_0^1 (\partial_x C_r) Q_r e^{\rho z} \, dx.$$

The derivative weights each count event by its number t of correct modal categories; the integral gives $t/(t + u)$ when u wrong categories tie. For diffuse wrong mass d, use e^{dxz} at $r = 1$ and e^{dz} at $r \geq 2$; abstention always contributes $e^{\rho z}$. Polynomial truncation at degree K and Gauss–Legendre integration evaluate the finite expression up to floating-point roundoff, without Poisson approximation. Independent ordered and rational count enumeration check ties, multiple correct labels, invalid abstention, extremizers, and nested information sets. Multiclass voting bounds are established prior work (Aeeneh et al., 2023); our use here is an information-ablation procedure for cross-system decisions.

Retrospective information ablation. We reuse only the development questions of Appendix W: 1,800 SuperGPQA and 92 math questions, with all 80 saved responses per system. No literal answer categories are merged using correctness. In particular, multiple correct categories occur on 1/6 SuperGPQA questions for Qwen/high and on 7/18/30/34 math questions for Qwen/low/medium/high. We retain them all. For each question, information level one retains total correct mass; level two retains its full correct-category vector and abstention mass; level three additionally retains maximum eligible wrong mass. We average per-question endpoints, never apply a nonlinear bound to mean correctness.

Across all 16 contexts (two tasks, two candidate sets, $K = 32, 64$, and two validity rules), the first two information levels never separate a unique winner. The third separates Qwen among all four systems and high effort among the three effort settings. Complete results, code, and the fixed protocol are released. These decisions agree with pass@1 and direct evaluation; the experiment establishes information narrowing, not a better selector or sample-efficiency advantage. Finite-response estimation error remains outside the identification ranges, and a question bootstrap alone would not account for unseen answer categories. The question-disjoint test in Appendix W evaluates the resulting effort choice, but does not establish superiority of this particular screening procedure.

X.1 Decision-preserving verifier queries

Once a selector is specified, must every candidate be scored to execute it? For bounded weighted voting, the answer is no. We implement a fixed-pool specialization of classical bound-based aggregation (Fagin et al., 2002). The goal is to reproduce full scoring, not to improve its answer or certify correctness. Unlike a confidence threshold for stopping generation, this procedure retains every generated response and skips only score queries.

Access model and algorithm. The mapper supplies all eligible category labels before scoring. Each response has a fixed weight $w_i \in [0, 1]$, available only by querying its scorer. Category sums determine the winner; positive ties follow first occurrence. If all sums vanish, the result is the plurality fallback f. For the observed weights, write L_a for category a's partial sum and n_a for its

Table 42: Fixed-score query replay at $K = 64$, structured-semantic mapping. Query counts are means over all 500 questions. Token saving compares the largest-upper order with eligible-only full scoring. The hindsight row is an offline unit-query lower bound, not an executable baseline.

Query rule or diagnostic	Base	GRPO-42
Full eligible scoring	42.630	46.484
Source-order certificate	19.254	19.178
Random-order certificate	19.180	19.454
Largest-upper certificate	14.038	14.014
Hindsight minimum	11.124	10.990
Scorer-input tokens saved (%)	66.06	67.81
Preserved selected accuracy (%)	75.4	75.8

unqueried response count. Its full sum lies in $[L_a, U_a]$, where $U_a = L_a + n_a$. Return a as soon as its positive lower bound beats every competing upper bound, allowing equality only when a has earlier tie priority. Alternatively, return f when $U_b = 0$ for every $b \neq f$. Otherwise query the earliest remaining response in the category with largest U_a and repeat, breaking category ties by first occurrence. No eligible category means abstention; a sole eligible category needs no query. Neither correctness labels nor unrevealed scores enter this interface.

Proposition 6 (Complete fixed-score stopping certificate)**.** *For a nonempty eligible category set, let each unqueried weight vary freely over* $[0, 1]$*. The observations determine the full weighted-vote winner if and only if one of the two stopping conditions above holds. Therefore the algorithm returns exactly the full-scoring answer after at most one query per eligible response. Sufficiency also holds when hidden scores are dependent.*

Proof. For sufficiency, the full sums lie in their stated intervals, so a separated positive lower bound wins under every completion. If only f can have positive mass, it wins whether its sum is positive or all sums are zero. For necessity, first suppose some $L_a > 0$. Setting every unobserved weight to zero gives the tie-broken maximizer a of L. If its positive certificate fails, a competitor can attain its upper bound while a remains at L_a, defeating a strictly or by tie priority. These are feasible completions with different winners. If every L is zero, the zero completion gives f; any non-fallback category with positive U can instead win by assigning it a positive hidden weight. Thus the second condition is necessary. Each nonterminal step reveals one previously unknown weight, establishing termination. □

Minimum witness with hindsight. The certificate also gives an exact unit-query lower bound for a realized score vector. Let a be its full winner. Sort its scores descending as $x_1, \ldots, x_{N_a}$ and each competitor's scores ascending as $y_{b,1}, \ldots, y_{b,N_b}$. With h winner queries, the largest attainable lower bound is $A(h) = \sum_{j \leq h} x_j$; with q competitor queries, its smallest attainable upper bound is $B_b(q) = N_b - \sum_{j \leq q}(1 - y_{b,j})$. For each h with $A(h) > 0$, take the least $q_b(h)$ making $A(h)$ beat $B_b(q_b)$ under the tie rule, and minimize $h + \sum_{b \neq a} q_b(h)$ over feasible h. If $a = f$ and all competing scores are zero, also consider querying every response outside f and none inside it. For fixed per-category query counts, these prefixes maximize the winner's lower bound and minimize each competitor's upper bound independently. Together with Proposition 6, this proves the minimum witness size. The witness uses hidden scores and is an offline diagnostic, not a deployable query strategy. Hindsight-witness comparisons also arise in active preference elicitation (Shah & Yu, 2026); our score-access model differs, and no worst-case competitive ratio or online-optimal ordering is claimed.

Complete saved-score replay. We retain both policies, all 500 MATH-500 questions, three mappers, $K \in \{8, 16, 32, 64\}$, and three query orders: largest-upper, source order, and one fixed random order. All 36,000 selections agree with full exact scoring and the published stable log-product implementation. The certificate uses exact rational products of saved floating-point step probabilities; empty step vectors have zero weight. Every queried response's recorded scorer-input length is charged, including queries that do not improve the partial winner. The reference scores only eligible responses, so savings exclude the trivial benefit of skipping invalid answers.

Table 43: Batch-preserving verifier-stage time on 64 questions per policy. Entries are lazy/full time ratios, with one-sided, six-comparison-adjusted 95% bootstrap upper bounds in brackets. Smaller is better; 1 means no saving. Both methods use the same partition.

Fixed scoring partition	Base	GRPO-42
Singleton	0.312 [0.450]	0.304 [0.437]
Source-order blocks of four	0.487 [0.609]	0.424 [0.547]
Length-sorted blocks of four	0.372 [0.507]	0.358 [0.491]

Largest-upper uses 68.8/69.7% of source-order certificate tokens and 68.8/69.3% of random-order tokens on Base/GRPO. Its total query count is 1.262/1.275 times the hindsight minimum; these are dataset-level ratios, not per-instance guarantees. The complete witness audit checks all 12,000 policy–mapper–budget–question records and every prefix of all three query traces. Full reports and CPU replay programs are released.

Batch-preserving execution. All candidate generation is still required. Live batching requires additional care: changing a response's batch can change its numerical score. An implementation may query whole predeclared blocks and apply the same certificate after assimilating each block, but exact preservation then assumes each block's score vector is invariant to which other blocks have been queried. Fixed membership alone does not prove deterministic GPU execution. The unit-query hindsight minimum does not optimize block count or hardware time.

We measure this implementation on 64 hash-selected MATH-500 questions per policy at $K = 64$, excluding both earlier timing subsets. Four L40S GPUs each handle 16 questions; each full/lazy pair runs on the same device, with two repetitions and rotated execution order. Before scoring, we fix three eligible-only partitions: singleton blocks, source-order blocks of four, and length-sorted blocks of four. At each step, we choose the category with the largest upper bound among those with unqueried responses, breaking ties by first occurrence. We query the fixed block containing its earliest unqueried response and incorporate every score in that block before checking the certificate. Lazy execution never regroups the remaining candidates. The timed region includes tokenization, forward passes, exact aggregation, certificate checks, and the common planning cost. It excludes model loading, source I/O, and previously completed candidate generation.

All 1,536 trials are retained and independently replayed. Within each fixed partition, every queried weight and selected label agrees with full scoring. Table 43 reports all six configurations: verifier-stage time falls by 51.3–69.6%. Bounds use 10,000 paired question bootstraps within GPU shards after averaging the two repetitions; they are approximate bootstrap bounds, not finite-sample guarantees. The question set and the requirement that all six ratios be at most 0.8 with upper bounds below 1 were fixed before these timings. Existing response banks were reused, so this is new hardware evidence, not an independent accuracy confirmation.

An exact retrospective replay holds these partitions and scores fixed while varying query order. Both comparison rules use the same stopping certificate: one follows the predeclared block order, and the other uses a fixed random permutation. With blocks of four, largest-upper requires 4.4–7.1% fewer block queries than predeclared order and 8.9–15.1% fewer than random order across both policies and partitions. For singleton blocks, the corresponding ranges are 27.1–31.3% and 28.4–30.2%. All 2,304 replayed selections agree with full scoring. These are query-count comparisons, not additional latency measurements; the released report retains every context and repetition.

Batch context remains material: scores differ across partitions even though their selected labels agree on these 64 questions. An earlier serial-versus-batched pilot changed one selected answer, motivating the fixed-partition comparison. Execution failures and complete-shard reruns are documented with the released timings; partial trials are not mixed into the aggregate. These measurements establish verifier-stage savings for this scorer and hardware, not end-to-end superiority, an optimal serving implementation, or a more accurate selector.

Y Extended Related Work

This section expands the main-text related work with the literature that most directly constrains how a residual selection gap should be read. It supplies context only; no reported result depends on it.

Coverage, modal limits, and system choice. Bay & Yearick (2026) separate coverage from plurality selection and derive its unique-mode limit, alongside a distinct analysis of benchmark-estimation correlation. Romano et al. (2026) study the gap between candidate quality and its recovery on open-ended generation tasks. These establish the broader selection bottleneck. Our oracle-curve construction asks which system an evaluator would choose, including strict oracle dominance with limiting plurality regret one. The empirical audit then distinguishes pairwise crossings from an overall-winner change. The disjoint categorical predictor targets finite-budget residuals under a fixed mapper and voting rule.

Zero-RL recipes and evaluation sensitivity. RL-for-reasoning work addresses objective design, systems implementation, process rewards, and entropy or diversity control (Shao et al., 2024; Yu et al., 2025; Liu et al., 2025b; Fu et al., 2025; Cui et al., 2025a; Jiang et al., 2025; Lochab et al., 2026). Open replications of R1-style training show that verifier-only rewards raise mathematical pass@1 across many open base models, and that the size of the reported gain depends on prompt template, data filtering, and length control (Zeng et al., 2025; Hu et al., 2025). Our canonical configuration follows that lesson on the training side, and the answer-mapper sensitivity in Table 1 is its selection-side analogue: the residual survives every mapper, whereas the training contrast does not.

Entropy, diversity, and where probability mass moves. A growing line of work treats policy entropy as the resource RLVR consumes, measuring its collapse (Cui et al., 2025b), attributing updates to a high-entropy minority of tokens (Wang et al., 2025), or regularizing it explicitly (Jiang et al., 2025; Lochab et al., 2026). Aphale & Liu (2026) show a different rank inversion: deeper SFT can improve pre-RL pass@1 yet worsen the checkpoint's post-RL pass@10 through entropy and reward-variance collapse. Our systems are already trained; their ranking changes only when oracle coverage is replaced by the deployed selector. Our measurements are consistent with entropy reduction but show that its sign alone is uninformative for selection: canonical GRPO lowers normalized entropy while moving mass toward the correct answer, and released ProRL lowers it sharply while improving every accuracy endpoint. What a plurality selector responds to is the modal margin Δ_j, not a scalar spread.

Process supervision and outcome verification. Process- and outcome-based feedback were compared directly by Uesato et al. (2022). Later work removed the human-annotation bottleneck with automatically constructed step labels (Wang et al., 2024b; Luo et al., 2024), with implicit process rewards derived from outcome labels (Yuan et al., 2024), and with verifiers trained as next-token predictors (Zhang et al., 2024); Zheng et al. (2024) then showed that identifying the erroneous step remains hard. The public PRM results in Table 29 compare three aggregations after positive controls; product-weighted consensus improves plurality, but is not a bound on selection. Weaver instead combines weak verifiers and distills their ensemble to shrink the generation–verification gap (Saad-Falcon et al., 2025). Our selector spectrum is complementary: it measures multiple named selectors and asks whether oracle evaluation changes the choice among trained systems, rather than proposing another verifier.

Training for the deployed inference rule. Tang et al. (2025) optimize multi-sample objectives, including majority voting, using a leave-one-out policy-gradient estimator. They also characterize how centering its credits changes the optimized objective. Ousherovitch & Tewari (2026) derive strategy-dependent training weights for coverage, voting, and best-of-N. Thus neither the general training–inference mismatch nor direct optimization of voting is new here. Our contribution concerns what oracle curves fail to identify about system choice, and the empirical conditions under which that distinction changes rankings. The unsuccessful SAPO probe tests one heuristic intervention; it does not establish that inference-aware training is ineffective.

Allocating and routing test-time compute. Setlur et al. (2025) argue that scaling test-time compute without verification or RL is suboptimal, Sareen et al. (2025) unify a reasoner with a learned

verifier to improve test-time scaling, and Lin et al. (2026) route between voting and rewriting according to sample disagreement. Our frozen Gen-7B selector is a deliberately simple, non-adaptive instance of this family: it solves each problem once and matches its answer to an existing cluster. Under structured-semantic mapping it gains 9.2/7.0 points over Base/GRPO plurality while leaving 9.6/9.8-point oracle residuals. The gains largely reflect standalone solver capability: the pool adds only 0.6/1.2 points, with simultaneous intervals spanning zero. Thus this comparison does not establish complementary value from the pool or a compute-efficiency advantage.

Pass@k inference regret. Di et al. (2026) study a different deployment contract: a fixed reference generator samples N responses and may return up to k distinct candidates. They prove constant worst-case regret for frequency-based selection and derive a minimax-optimal Best-of-Majority rule under reward-model assumptions. Our oracle-induced regret instead concerns model choice: an evaluator uses each trained system's oracle coverage to choose one system, after which a named selector must return one answer. Proposition 1 isolates the resulting cross-system reversal even under strict oracle dominance at every budget; the uniform ranking audit then tests whether this possibility changes choices among observed RLVR and matched-RFT systems. The two analyses are complementary, and we do not claim a new within-generator minimax selector.

Consistency and calibration. Self-consistency both improves accuracy (Wang et al., 2023) and sharpens confidence estimates (Wang et al., 2024a). Our disjoint-pool analysis separates the two: Pool A supplies a problem-level gap risk whose aggregate calibration is within one point on Pool B, while its per-problem mean absolute error stays above eight points. Good aggregate calibration therefore does not license per-problem claims, and we report both.

Z Limitations of the Appendix Analyses

The full-parameter primary comparison uses 1.5B models; both 7B experiments are LoRA boundaries and do not isolate parameter count or architecture from update geometry. Three seeds are meaningful but not sufficient to resolve very small method differences. The Mistral training boundary covers one additional architecture; the public Qwen3.6/gpt-oss banks compare already-trained configurations, not training effects. The initial MBPP diagnostic uses math-trained policies; the separate DeepCoder-14B comparison adds a code-RL checkpoint and its same-backbone initializer, but is not a compute-matched causal control. Its 5.19% cap rate and possible pretraining contamination qualify this transfer evidence. MATH-500 confidence intervals characterize this benchmark, not an infinite task population. The disjoint pools are independent responses within a saved generation, while repeated partitions reuse samples. Answer normalization does not prove symbolic equivalence. The compatibility fallback can group malformed last lines; complete-box voting can discard valid unboxed answers; and the structured semantic grammar covers only exact rationals and tuples. Method rankings are consequently not claimed to be mapper-invariant. RFT is a matched-pass@1 reference intervention, not a causal control: compute, accepted data, objective, and optimization path differ. The public PRM can be miscalibrated under policy shift. ProRL uses a distinct decoding protocol. AIME has low power. Finally, selection gap is always relative to a selector; oracle coverage is not a realizable deployment system, and no tested selector establishes an upper bound on all future verifiers.