THE HIDDEN ALLOWANCE AXIS IN TEST-TIME COMPUTE COMPARISONS

Anonymous authors
Paper under double-blind review

ABSTRACT

A parallel test-time compute baseline chooses both sample count and per-sample token allowance. We show that fixing the allowance can conceal stronger baselines and change strategy comparisons. Two studies each select configurations on 200 validation questions and evaluate them on 400 held-out questions with three new test seeds. On MATH level 5 with two models, freeing the allowance at fixed sample count raises pooled accuracy by 7.58 points, at 23.4% higher input-plus-output token cost. A preregistered extension to four non-mathematical MMLU-Pro categories and three models opens the same allowance/count search to natural and reconsideration-suffix sampling. At the primary 32768-token selection target, validation-selected mixtures of natural configurations improve pooled accuracy by 1.76 points at 16.86% lower cost than the jointly tuned suffix, meeting the registered accuracy/cost criterion. The other primary comparison, against a fixed-allowance baseline, reduces cost by 19.01% without establishing accuracy superiority; single-configuration selection shows different cost trade-offs. Six historical grids quantify allowance sensitivity, and a plurality-vote account predicts aggregate accuracy with 2.2-point mean absolute error on held-out samples of the same questions. Natural-sampling controls separate prompt-associated discordance from ordinary variation. These results support fairer test-time strategy comparisons that jointly select both allocation controls and report measured costs.

1 INTRODUCTION

A reasoning model can spend additional inference compute on more attempts or longer attempts. Self-consistency samples independent reasoning paths and votes over their answers (Wang et al., 2023); sequential strategies extend or revise a reasoning process (Muennighoff et al., 2025; Snell et al., 2025). **A parallel baseline also has a depth control: the token allowance of each sample.** Fixing it selects one slice of a larger design space. How much of a strategy's apparent advantage depends on that slice?

We answer this question with an independently selected comparison. Configurations are frozen on 200 validation questions and evaluated on 400 held-out MATH level-5 questions, using two pinned model revisions and three new seeds. At the primary validation selection target of 32768 tokens, freeing the parallel allowance raises accuracy from 66.42% to 74.00%. Both policies select four samples: the **7.58-point improvement comes from a larger allowance at the same sample count**, with 23.4% higher actual input-plus-output cost.

The complete comparison reveals why accuracy and cost must be reported together (Figure 1). Joint allowance/count selection exceeds the fixed-eight-sample reconsideration-suffix family by 17.67 points while consuming 5.60% fewer tokens in the pooled comparison. A single long solve and adapted Self-Refine are less accurate on average but cheaper. All four accuracy contrasts are positive; one of four meets the registered accuracy/cost criterion. The nominal selection target alone would conceal these different realized expenditures.

The comparison also holds against a jointly tuned suffix outside mathematics. A separately registered extension evaluates four MMLU-Pro categories with three models, giving both natural and suffix sampling the same allowance/count grid. At the primary target, validation-selected mixtures of natural configurations gain 1.76 points at 16.86% lower pooled cost than the suffix. Of its two

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.

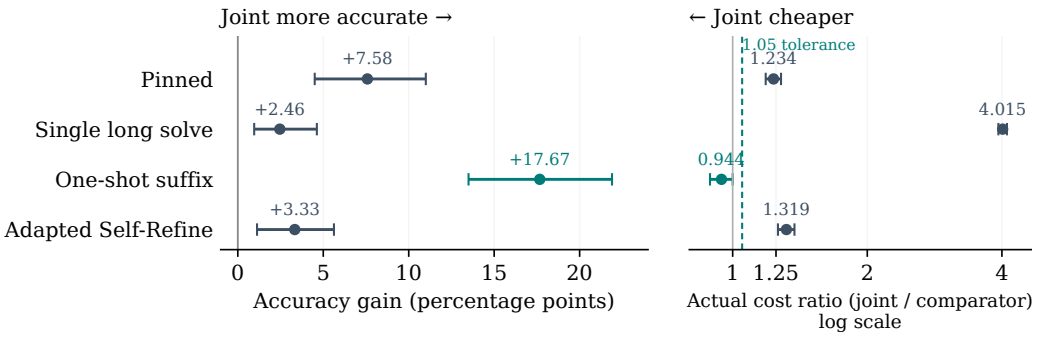

Figure 1: **Expose both the gain and its price.** All four registered MATH level-5 comparisons at the primary selection target. Points compare joint allowance/count selection with each baseline; bars are adjusted question/seed bootstrap envelopes. The accuracy lower bound must exceed zero and the cost-ratio upper bound must be at most 1.05 to meet the joint criterion. Only the suffix comparison meets both. Table 1 provides absolute results, settings and every interval.

primary comparisons, this one meets the joint criterion; against the fixed-allowance baseline, cost falls 19.01% without establishing accuracy superiority (Section 3.2).

The fixed allowance is a concrete experimental choice. For example, Sharma & Chopra (2025) compare independent chains with iterative refinement under a schedule that fixes each chain or step at 4096 tokens. Our earlier protocol similarly fixed the parallel allowance at 8192 while increasing the allowance of an arm with a one-shot suffix. We audit that allocation choice directly, rather than treating either fixed value as the natural definition of parallel sampling.

The historical grids reveal its magnitude and mechanism. Across six model–task grids, raising the allowance from 2048 to 8192 at 64 samples produces 2.0–6.3 times the gain of raising the sample count from 1 to 64 at allowance 8192. On Qwen3-8B/AIME, expanding the natural baseline from the pinned row to the full grid changes the suffix's apparent lead from 30.0 to 1.7 points. A standard voting-limit account explains why the controls need not substitute: more samples concentrate votes around the current answer distribution, while a larger allowance can change which answer is modal. An empirical-distribution predictor tests this account on held-out samples at the same questions and allowances. A separate natural-versus-natural control calibrates correctness discordance against ordinary sampling variation.

Contributions.

1. **Independently evaluated allocation comparisons.** Separate validation/test studies on mathematics and non-mathematical multiple choice evaluate five families, including iterative self-feedback. The main text reports every primary comparison; the second study gives both natural and suffix sampling the same search space (Section 3).
2. **Magnitude and a distributional account.** Six parallel grids, including a 32B model, quantify allowance sensitivity. An empirical voting model gives 2.2-point mean absolute error in same-question finite-sample calibration, with three descriptive baselines (Section 4).
3. **A controlled attribution protocol.** Matched-prompt and natural-sampling comparisons separate allocation, prompting and ordinary variation (Section 5), yielding a concrete evaluation protocol (Section 6).

2 Setup

We distinguish three controls: the sample count N, the maximum generated-token allowance c per sample, and the prompt. The historical natural-prompt grid uses $N \in \{1, 2, 4, 8, 16, 32, 64\}$ and $c \in \{2048, 4096, 8192, 16384\}$. We aggregate extracted answers by plurality vote, conventionally called majority voting in self-consistency. Invalid answers abstain; an all-invalid vote is incorrect.

Ties follow the first occurrence in sample order in the historical and MATH studies; the MMLU-Pro extension uses a frozen question/label hash shared across strategies.

One-shot reconsideration. Our suffix arm appends `\n\nWait, let me reconsider...` to the prompt and generates each sample in one call. The historical study uses $N = 8$ and a seven-rung allowance ladder from 512 to 16384. This is a prompt intervention, distinct from iterative refinement or budget forcing, which condition or intervene on generated text. Allowance effects and voting-limit estimates in Section 4 use only natural-prompt generations. Historical cross-arm comparisons use this one-shot suffix. Section 3 additionally evaluates iterative self-feedback; Appendix J documents the separate forcing attempts.

What is matched, and what is measured. Prompt comparisons fix the question set, N, and c and vary the suffix. Lever comparisons change one named control. Grid maxima measure search-space sensitivity. Their configurations can consume different amounts of compute: Nc bounds the allocation, while early stopping determines actual consumption. Historical generation-length fields count *characters*, which we retain as the unit of length comparisons. The configuration-verified Qwen3-8B/MATH-L5 rerun additionally records actual output token counts and stopping reasons (Appendix L.2). The natural-grid banks lack corresponding token records. The independent studies in Section 3 record actual input and output tokens for every strategy, including repeated feedback context.

Models, tasks, and coverage. The campaign includes six reasoning models: five DeepSeek-R1-Distill checkpoints at 1.5B, 7B, 8B, 14B and 32B, and Qwen3-8B (Guo et al., 2025; Yang et al., 2025). The five historical task families are AIME, MATH level 5, MATH-500 (Hendrycks et al., 2021; Lightman et al., 2024), GPQA Diamond (Rein et al., 2024), and BIG-Bench Hard ruin-names (Suzgun et al., 2023). Table 3 contains six natural-prompt grids, including the historical 32B sweep; the historical matched-suffix analysis contains sixteen cells on the other five models. All five cross-arm grid comparisons use 60 common questions. The historical matched-suffix cells retain their original intersections of 31–100 questions, a separate cohort from the later runs.

Appendix A specifies decoding, extraction and subsets. Appendix Q retains the prospective tests of cheaper allocation diagnostics.

3 Independent confirmation with measured token costs

3.1 MATH level 5: isolating the allowance effect

We freeze 200 validation and 400 held-out MATH level-5 questions, excluding exact canonical matches to the historical manifests. Qwen3-8B and DeepSeek-R1-Distill-Qwen-7B use pinned revisions and three validation seeds; the test uses three different seeds. All selected settings are locked before test generation. The evaluation has 400 independent question units, with models and seeds paired within question.

Five strategy families and a fixed decision rule. *Joint* selects both natural-prompt allowance and sample count; *pinned* fixes $c = 8192$ and selects N; *depth* fixes $N = 1$ and selects c. The *suffix* retains $N = 8$ and selects c. Adapted *Self-Refine* (Madaan et al., 2023) generates a solution, critiques it, and revises it, selecting the solution allowance, trajectory count and one or two rounds. The validation search covers 44 distinct configurations per model: 24 natural, four suffix and sixteen Self-Refine, with pinned and depth selecting subsets of the natural grid. Natural sampling uses $N \in \{1, 2, 4, 8, 16, 32\}$. Natural and suffix allowances span 2048–16384; Self-Refine uses 4096 or 8192, with $N \in \{1, 2, 4, 8\}$ trajectories. The suffix family's fixed $N = 8$ and these registered search spaces define the strategy comparisons. Selection maximizes validation accuracy subject to mean actual input-plus-output cost, breaking ties by lower cost and then configuration ID. The primary selection target is 32768 tokens; 16384 and 65536 are descriptive secondary targets. A target is a validation mean-cost threshold, not a per-question cap or equality of test costs. Each policy is charged for every logical call, including repeated feedback context (Appendix M). For Self-Refine, N counts trajectories and c caps each solution/revision call; feedback is capped at 2048 tokens. Both models select one round.

Table 1: **MATH level-5 results and every registered primary comparison.** The 32768-token target constrains validation mean cost. Test means average two models, three seeds and 400 questions. Both models select the same settings. Costs include all logical input and output tokens. Intervals are adjusted question/seed bootstrap envelopes; decisions use unrounded endpoints.

A. Selected policies and absolute performance

Strategy	c	N	Accuracy (%)	Tokens/question
Joint	16384	4	74.00	25,720
Pinned	8192	4	66.42	20,840
Single long solve	16384	1	71.54	6,405
One-shot suffix	4096	8	56.33	27,245
Adapted Self-Refine (one round)	8192	1	70.67	19,503

B. Joint minus comparator: gain, price and decision

Comparator	Δ accuracy [interval], pp	Cost ratio [interval]	Both criteria?
Pinned	$+7.58$ $[+4.50, +11.00]$	1.234 $[1.185, 1.283]$	No (cost)
Single long solve	$+2.46$ $[+0.96, +4.62]$	4.015 $[3.928, 4.105]$	No (cost)
One-shot suffix	$+17.67$ $[+13.50, +21.90]$	0.944 $[0.890, 1.000]$	**Yes**
Adapted Self-Refine	$+3.33$ $[+1.12, +5.63]$	1.319 $[1.263, 1.375]$	No (cost)

Freeing the allowance improves held-out accuracy. Both models select $(c, N) = (16384, 4)$ for joint and $(8192, 4)$ for pinned. The **7.58-point gain holds the sample count fixed**; actual cost rises by 23.4%. The single-long-solve baseline reaches 71.54% at 6405 tokens per question. Increasing its count from one to four adds 2.46 points at $4.02\times$ the cost, connecting the allowance effect to diminishing returns from additional samples.

All four accuracy/cost comparisons. The registered joint criterion requires an accuracy-difference lower bound above zero and a cost-ratio upper bound at most 1.05. We construct 99.375% marginal intervals for four differences and four cost ratios using 20000 paired bootstrap draws. Each reported envelope covers both question-only and crossed question/seed resampling; Bonferroni adjustment targets a nominal 95% family level. Table 1 gives every interval and decision, conditional on the frozen policies. **One of four comparisons meets both criteria:** joint exceeds the suffix by 17.67 points at 5.60% lower mean cost. All four accuracy-difference intervals are positive. Against adapted Self-Refine, the 3.33-point gain costs 31.9% more; against pinned and depth, the cost increases are also outside the registered tolerance.

Joint exceeds the suffix in all six model/seed cells. Leaving out any one seed leaves all four pooled accuracy differences positive. Cost effects differ: the joint/suffix ratios are 1.039 for Qwen3 and 0.836 for Qwen-7B. Appendix M reports all seeds and secondary targets, including the infeasible low-target Qwen3 suffix and the higher-target ranking reversal. The selection target alone does not determine the ranking.

3.2 Non-mathematical extension with equally tuned sampling families

A separate preregistered study uses biology, history, law and philosophy from MMLU-Pro (Wang et al., 2024): 50 validation and 100 held-out questions per category, shared across Qwen3-8B, R1-Distill-Qwen-7B and R1-Distill-Llama-8B. Both natural and suffix families now search the same four allowances and six sample counts up to 32. With Self-Refine, the search has 64 configurations per model. The primary selector maximizes validation accuracy under the mean-token target, allowing mixtures of at most two complete configurations. A mixture represents a fixed, question-independent lottery over configurations, not a pooled vote across them. Weights are frozen before testing; reported accuracy and cost are exact weighted means of the component values. Single-configuration selection is a registered sensitivity check.

A stronger suffix comparison. At the primary 32768-token target, joint natural selection improves pooled accuracy by **1.76 points at 16.86% lower cost** than the jointly tuned suffix. Of two registered primary comparisons, this one meets the joint criterion; versus pinned, the difference is $+0.17$ points

Table 2: **MMLU-Pro extension: complete primary comparisons and secondary controls.** Joint natural accuracy is 44.73% at 27115 input-plus-output tokens/question. Means pool three models and fixed seeds over 400 shared questions at the 32768 validation target, using the registered mixture selector. Brackets are descriptive 95% intervals; primary decisions use the adjusted one-sided bounds stated in the text. Both primary comparisons are retained.

Comparator	Δ accuracy [95% CI], pp	Cost ratio [95% CI]	Primary gate
Pinned	$+0.17$ $[-0.60, +0.94]$	0.810 $[0.801, 0.819]$	No
Jointly tuned suffix	$+1.76$ $[+0.53, +3.01]$	0.831 $[0.818, 0.845]$	**Yes**
Adapted Self-Refine	$+1.84$ $[+0.78, +2.90]$	0.819 $[0.809, 0.829]$	Secondary
Single long solve	$+7.04$ $[+5.66, +8.44]$	9.686 $[9.465, 9.917]$	Secondary

at 19.01% lower cost without establishing accuracy superiority. The 20000-draw bootstrap resamples paired question clusters within categories, averaging seeds and weighting models and categories equally. Adjusted one-sided bounds use quantiles 0.0125 and 0.9875 for the four primary endpoints: the suffix comparison's accuracy lower bound is $+0.35$ points and cost-ratio upper bound 0.847; pinned's accuracy lower bound is -0.72 points.

What the selected policies reveal. Qwen3 selects $(c, N) = (16384, 4)$, while Llama mixes $(2048, 16)$ and $(4096, 32)$; Qwen-7B's joint and pinned mixtures coincide. Thus explicit allowance selection need not mean increasing the allowance. Joint's pooled advantage over the suffix has the same direction in each test seed, but its cost saving is mainly from Qwen3 and its accuracy gain mainly from Llama; Qwen-7B's accuracy difference is -1.06 points. Under the registered single-configuration selector, the pooled suffix comparison gains 1.61 points at 8.09% higher cost. Appendix N reports the complete model, seed, target and selector results.

A larger-model check. A separately registered R1-Distill-Qwen-32B study reuses the same question split and primary mixture selector. At 32768, joint and pinned select the same $(c, N) = (8192, 16)$ policy; joint exceeds the tuned suffix by 2.50 points at $3.36\times$ the measured cost. Neither primary comparison meets the joint criterion (Appendix O).

4 The per-sample allowance is a consequential lever

4.1 Separate allowance from prompting

A comparison that simultaneously adds a suffix and raises its generation allowance combines two interventions. The appropriate controls are a natural-prompt grid over allowance and sample count, and a prompt ablation at the same coordinates. These answer different questions: how much the baseline depends on allocation, and what changing the prompt contributes at a chosen allocation.

4.2 The allowance has a larger gain at the tested endpoints

Table 3 measures the two controls entirely within the parallel arm. **The allowance gain is larger in all six grids**, ranging from 31.7 to 46.7 percentage points, compared with 5.0 to 21.7 for the sample-count gain. The effect persists at 32B, with a ratio within the smaller-model range.

The ratio compares a fourfold allowance increase with a sixty-fourfold sample increase. A supplementary equal-multiplier comparison, starting at $N = 1$, finds a larger allowance gain in 35 of 36 tested cases, with both gains zero in the remaining case (Appendix I). This is supporting sensitivity analysis; the first-occurrence tie rule can make $N = 1 \to 2$ uninformative. At the main table's endpoints, fixing c leaves substantial accuracy available within the parallel baseline. The Qwen-14B main bank is selected by source provenance; an alternative complete bank preserves the grid maximum and gives an extra-depth gain of 8.3 rather than 10.0 points (Appendix L.1).

Table 3: Accuracy gains within the natural-prompt parallel arm. Allowance gain: $c = 2048 \to 8192$ at $N = 64$. Sample gain: $N = 1 \to 64$ at $c = 8192$. Extra depth: $c = 8192 \to 16384$ at $N = 64$. All gains are percentage points. Each grid uses the same 60 questions across its four allowances. Ratios compare these endpoints, not equal realized token expenditures.

Model	Task	Allowance gain	Sample gain	Ratio	Extra depth
Qwen-7B	AIME	+33.3	+11.7	2.86×	+15.0
Qwen-14B	AIME	+43.3	+21.7	2.00×	+10.0
Qwen3-8B	AIME	+38.3	+11.7	3.29×	+26.7
Qwen-32B	AIME	+46.7	+18.3	2.55×	+13.3
Qwen-7B	MATH level 5	+31.7	+5.0	6.33×	+1.7
Qwen3-8B	MATH level 5	+41.7	+6.7	6.25×	+10.0

Table 4: Same-question finite-sample calibration. Lower mean absolute error (MAE) is better. All predictors use only estimation samples 0–31; votes on 32–63 provide the observed accuracy. MAE averages 144 correlated cell/allowance/count points, with 60 questions per cell. The new constant baseline uses all 32 estimation samples; invalid answers count as incorrect.

Predictor	MAE (percentage points)
Empirical distribution	2.17
Constant: first sample	6.44
Constant: 32-sample mean	6.26
Replay vote@N	2.62

4.3 A DISTRIBUTIONAL ACCOUNT OF THE TWO CONTROLS

Fix a question, prompt, allowance and extraction procedure. Invalid outputs abstain. Under independent draws with positive valid-output probability and a unique most probable valid answer, plurality voting selects that answer with probability tending to one as N grows. Averaging over questions gives the limiting accuracy

$$L(c) = \Pr_{\text{question}}[\text{the unique modal extracted answer at allowance } c \text{ is correct}].$$

Here all-invalid questions have zero voting success; population ties require separate treatment. We use *ceiling* as shorthand for this asymptotic limit, not a finite-N upper bound. Finite-sample accuracy can exceed the limit, and convergence need not be monotone.

Increasing N samples the same distribution more often; changing c can change the distribution itself and thus $L(c)$. A prompt or decoding change can also move this limit. Width concentrates the vote around the current mode; depth can change which answer is modal. Whether this improves accuracy depends on the resulting distribution.

Same-question finite-sample calibration. For each question, we estimate the answer distribution from sample IDs 0–31, simulate voting from it, and evaluate against votes on IDs 32–63. The model uses the estimated answer frequencies directly, with no fitted scaling-law parameters. Table 4 compares it with three baselines, including a constant-in-N prediction that uses all 32 estimation samples. **The model's 2.17-point error is substantially below this 6.26-point constant baseline**, supporting the role of answer frequencies in explaining the finite-sample curve. Replay uses the first N estimation samples directly for vote@N, rather than simulating votes from all 32 samples' answer frequencies.

A paired bootstrap over 120 shared questions supports the improvement over the constant baseline; the smaller gain over replay depends on the interval form. The 144 grid points are not independent observations. This is calibration on new draws of the same questions, not prediction on unseen questions. Appendix H gives both banks and interval forms.

Table 5: Same-question baseline sensitivity. The suffix arm uses $N = 8$; the natural-prompt row and grid each select their highest observed accuracy. The *vs grid* column is the suffix peak minus the grid maximum, not a matched-cost difference. All five comparisons use 60 common questions. The 95% ranges are paired question bootstraps conditional on the selected configurations and recorded banks. Qwen-14B uses bank A; its alternative bank is in Appendix L.1.

Task	Model	n_q	Suffix peak	vs row	vs grid [95% range]
AIME	Qwen3-8B	60	71.7	$+30.0$	$+1.7$ $[-3.3, +6.7]$
AIME	Qwen-14B	60	73.3	$+15.0$	$+3.3$ $[-3.3, +10.0]$
AIME	Qwen-7B	60	55.0	$+8.3$	-6.7 $[-15.0, 0.0]$
MATH-L5	Qwen3-8B	60	76.7	$+8.3$	-1.7 $[-5.0, 0.0]$
MATH-L5	Qwen-7B	60	73.3	0.0	0.0 $[-6.7, +6.7]$

4.4 Expanding the baseline changes the apparent advantage

We compare the best suffix-arm rung with two natural-prompt baselines: the best point on the pinned $c = 8192$ row, and the best point on the full grid. **On Qwen3-8B/AIME, this changes the apparent lead from 30.0 to 1.7 points.** Table 5 gives every cell with the needed sweep.

Figure 2 in the appendix visualizes the Qwen3/AIME grid.

The MATH-L5 banks also agree at identical allocation coordinates. In the configuration-verified Qwen3-8B rerun, the suffix reaches 46/60 correct, compared with 41/60 for the pinned-row maximum and 47/60 for the grid maximum. At the suffix peak's exact coordinates, $N = 8$ and $c = 16384$, the natural arm also scores 46/60. The lost row-relative advantage therefore does not require increasing the baseline's sample count. The grid-relative difference itself is only one question; Appendix L.2 provides the complete ladder and configuration provenance. The repaired suffix has pinned weights, whereas the historical natural revision is unverified. This cross-bank comparison establishes allocation-coordinate agreement; the prompt ablations below use the historical matched-condition cells.

The grid offers 28 candidates against seven suffix rungs, so maxima can benefit unequally from selection on the evaluation questions. In 400 half-split checks that select on one half and score on the other, the mean grid-relative margins are $+3.8$, $+6.2$, and -6.8 for AIME, and -0.5 and $+0.4$ for MATH-L5, in table order. These selection-sensitivity results retain a substantial allowance effect while showing that residual margins depend on the selection procedure. The resulting attribution lesson is concrete: a lead against one pinned parallel row combines the choice of baseline allowance with the prompt intervention.

Where additional allowance stops paying. The effect is task-dependent. Doubling the allowance from 8192 to 16384 buys Qwen-7B 15.0 points on AIME but only 1.7 on MATH-L5; Qwen3-8B still gains 10.0 on MATH-L5. These differences make the model–task cell the useful unit of allowance calibration.

5 Separating prompt effects from sampling variation

Separating the allowance leaves a well-defined prompt intervention: generate with or without the one-shot suffix at fixed N and c. We distinguish its effect on vote accuracy from disagreement between the two sets of sampled outcomes.

5.1 Accuracy effects vary across tasks

Across the sixteen matched cells, the suffix has mixed vote-accuracy effects. Task-family means are positive for the three mathematics families and negative for GPQA and BBH (Table 6). For the fixed tested cells, the AIME interval is positive and the GPQA interval negative. The AIME result is also checked across four allowances. These contrasts show heterogeneous prompt effects across the tested model–task cells at a common allocation.

Table 6: Matched suffix comparisons and their natural-sampling control at $c = 8192$, using eight samples per group. n_q counts question–cell observations; discordance pools matched pairs and vote changes pool questions across the family's cells. D_0 and D_1 are percentages; E and vote changes are percentage points. Vote intervals hold cells fixed and resample shared question IDs jointly across models, preserving coverage and conditioning on recorded banks. Per-cell excess intervals appear in Appendix L.3.

Task family	Cells	n_q	D_0	D_1	E	Δ vote [95% CI]
MATH-500	2	200	7.1	6.4	-0.7	$+1.0$ $[-1.0, +3.0]$
MATH level 5	4	200	11.8	12.4	$+0.6$	$+1.0$ $[-1.5, +3.5]$
AIME	5	272	15.7	17.4	$+1.7$	$+7.0$ $[+4.0, +10.3]$
GPQA Diamond	3	112	32.5	33.8	$+1.3$	-8.0 $[-15.2, -1.8]$
BBH ruin-names	2	200	32.1	34.6	$+2.4$	-1.5 $[-7.0, +4.0]$

5.2 A NATURAL-SAMPLING CONTROL CHANGES THE INTERPRETATION

Let D_1 be the fraction of natural/suffix sample pairs whose correctness indicators disagree, pairing by question and sample index. This diagnostic includes ordinary sampling variation. To measure that component, we split the natural bank into samples 0–7 and 8–15 and pair the two groups on the same questions, obtaining natural–natural discordance D_0. The excess $E = D_1 - D_0$ measures excess correctness discordance relative to this no-suffix control. All comparisons retain the original sixteen source intersections. These are paired outcomes from separate draws, not revisions within a trajectory; disagreement between two incorrect answers is not counted.

Ordinary sampling already separates the observed task groups. Natural-only D_0 spans 5.8–22.3% across eleven open-ended cells and 28.2–35.1% across five multiple-choice cells, with no overlap. Raw natural/suffix discordance D_1 also separates them, at 5.0–24.2% and 31.1–38.0%, respectively. After subtracting the control, the excess ranges overlap: -1.0 to $+4.0$ points for open-ended tasks and -1.0 to $+6.0$ for multiple choice. Fifteen of sixteen cellwise conditional excess intervals include zero. The format-associated separation is thus already a property of the natural sampling banks, rather than a distinctive signature of the suffix.

This is a discordance diagnostic, not an equivalence test for prompt effects; the vote-accuracy contrasts measure the latter directly.

Supporting interventions. Alternative suffix wordings preserve similar raw discordance but can reverse the directional imbalance. Re-rendering the same mathematics questions as four-way multiple choice increases raw discordance in all four comparisons, although only one of six registered threshold predictions holds. The intervention also changes answer cues and effective difficulty. Appendix P retains the complete cohorts, distractor audit, directional comparisons and length diagnostics. The natural-sampling control in Table 6 identifies the additional measurement needed to attribute such variation to the prompt.

6 RELATED WORK, LIMITATIONS, AND IMPLICATIONS

Relation to prior work. Self-consistency (Wang et al., 2023), repeated-sampling studies (Brown et al., 2024) and inference-scaling analyses (Wu et al., 2025; Snell et al., 2025) establish that inference allocation and aggregation affect performance. Analyses of parallel–sequential gaps (Gu et al., 2026) and iterative-refinement comparisons (Sharma & Chopra, 2025) condition later attempts on earlier solutions. Bilal et al. (2026) combine independent initial paths with iterative self-critique and correction, then vote over the refined answers. Budget forcing (Muennighoff et al., 2025) actively controls continuation. Our historical suffix isolates a simpler prompt intervention; the independent studies additionally adapt the generate/feedback/revise mechanism of Madaan et al. (2023) to the same reasoning models and questions. Huang et al. (2024) identify evaluation pitfalls in reasoning self-correction, including weak baselines and unequal inference budgets. Our audit isolates per-sample allowance as a baseline-selection dimension. Allocating compute across questions and

choosing (N, c) within a parallel baseline are complementary decisions; our experiments address the latter.

Hariri et al. (2026) distinguish inference regimes, proposal distributions, evidence extraction, aggregation and compute accounting. Their effort-level analysis also identifies censoring at a shared output limit. Our natural-prompt grids vary the allowance directly and measure its effect on the baseline. A related audit by Dominguez-Olmedo et al. (2025) shows how task-specific training can confound model comparisons; our intervention instead concerns inference allocation at fixed model weights. Liao et al. (2026) vary reasoning-prefix depth and sampling width, aggregating answers from intermediate prefixes. We instead vary the cap on complete natural samples. Zhou et al. (2026) study correctness changes within reasoning trajectories. Our contribution quantifies allowance effects in natural sampling, combines matched-condition suffix tests with a natural-sampling control, and confirms selected strategies on separate questions with measured costs.

Limits of the evidence. Historical coverage stops at 32B and is not factorial. Cross-arm grids use 60 questions each and prompt cells contain 31–100; their conditional intervals omit full generation-seed and selection variability. Studies using separate validation and test splits cover MATH level 5 and four MMLU-Pro categories on 7B/8B models, with a separately registered 32B check on the same MMLU-Pro cohort. Their intervals condition on locked policies; the extension also conditions on its realized seeds, with the reported selector and model heterogeneity defining the pooled result's scope. Exact task-manifest exclusion does not establish model-training decontamination. Costs are measured in tokens, not latency or throughput: longer decodes add serial work while concurrent samples use additional memory, with rankings dependent on hardware and scheduling. Historical forcing gates and extraction coverage are documented in Appendices J and C.

An allowance-aware evaluation protocol. The findings support four concrete reporting steps. First, expose N and c rather than fixing one silently. Second, ablate prompting at identical (N, c) on the same questions and pinned model revision, including a natural–natural sampling control. Third, select configurations on validation questions and report held-out accuracy alongside actual token costs, reporting latency separately if relevant. Fourth, keep extraction failures, invalid tests and incomplete coverage visible. Together these steps distinguish the choice of a baseline, a prompt intervention, and ordinary sampling instability. Earlier cheap allocation diagnostics did not replace these measurements: the complete registered record and predictive baselines remain in Appendix Q.

Conclusion. A parallel baseline specifies both an answer distribution and a sample count. In the six grids studied here, changing its per-sample allowance produces large gains without a reconsideration prompt; on Qwen3-8B/AIME, changing the baseline search space reduces an apparent 30.0-point lead to 1.7. A standard voting-limit account explains why allowance and sample count need not substitute for each other. Independent MATH confirmation gives a 7.58-point gain from freeing the allowance at fixed sample count, with 23.4% higher cost. The non-mathematical extension gives both natural and suffix sampling the same search space: the registered mixture selector gains 1.76 points at 16.86% lower pooled cost than the jointly tuned suffix. Natural-sampling controls calibrate excess prompt discordance. Together, these findings turn a hidden baseline choice into an explicit axis of an independently evaluated accuracy/cost comparison.

Reproducibility statement

Appendix A specifies the historical generations and extraction; Appendix C gives the statistical readings, and Appendix D records corrections. The complete historical scoreboard is in Appendix Q, with available registration excerpts in Appendix R. Appendix M specifies the independent study's split, strategies, selection and cost rule.

The existing anonymous CPU package reproduces the historical and MATH evidence from a clean extraction, using Python and NumPy without model weights, a GPU or external data. The historical component retains its source-bank choices, question intersections and sample IDs, regenerating the grid, prompt-control and distribution-model results. The added calibration component reconstructs the 32-sample constant baseline and paired question-bootstrap intervals from the same frozen grids, retaining both alternative banks and interval forms. It also includes the 160 rendered option sets and their paired distinct-option audit. Its README lists analyses not reconstructed by the package.

For the independent confirmation, the package reconstructs all 30 selection slots from validation configuration summaries, all 34800 policy votes and logical costs from 110400 slim call records, and every primary bootstrap interval and decision. The 400 questions remain the independent units. Two implementations cross-check the primary intervals. Exact-answer and empty-vote grades are checked independently; other equivalence grades use the authenticated returned labels. Full symbolic regrading, inference and the full validation generation bank are outside this numerical release.

Distributed files have checksums and explicit source-to-anonymous-copy provenance. The local raw-subset audit checks token-array lengths against slim counters; the package documents its deterministic selection rule and audit result, but does not distribute the raw subset or full traces. The root command runs the historical, confirmation and calibration components and reports their coverage separately.

For the MMLU-Pro extension, a separate score-level CPU package contains the frozen contract, validation and test configuration scores, selection records and statistical tools. It reproduces all 90 selection slots and both primary decisions from 115200 validation and 40800 test score rows, using 20000 bootstrap draws. Appendix N specifies the protocol and coverage. This numerical reproduction does not independently regrade raw generations or reconstruct allocated GPU time from configuration scores. A separate 32B score package reproduces all 30 selections and both primary decisions from 38400 validation and 12000 test score rows (Appendix O), without pooling with the three-model extension.

AI USE STATEMENT

Generative AI tools provided limited assistance with hypothesis refinement. Generative AI tools assisted with experimental design, dataset preparation, conceptual and mathematical analysis, implementation and debugging of experimental and analysis code. They also supported literature search and manuscript revision and editing. Reported measurements were derived from recorded experimental outputs. Verification included configuration audits, automated tests, numerical reproduction, and checks against primary sources.

REFERENCES

Ahsan Bilal, Muhammad Ahmed Mohsin, Muhammad Umer, Lena Trigg, Ali Subhan, Muhammad Ali, and Dean F. Hougen. Refining over resampling: Test-time self-correction for LLM reasoning, 2026. URL https://arxiv.org/abs/2608.05643. arXiv preprint arXiv:2608.05643.

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. URL https://arxiv.org/abs/2407.21787. arXiv preprint arXiv:2407.21787.

Ricardo Dominguez-Olmedo, Florian Eddie Dorner, and Moritz Hardt. Training on the test task confounds evaluation and emergence. In *International Conference on Learning Representations*, pp. 68564–68594, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/ab8c971c2ccd12bac0ab249f75e2c16d-Paper-Conference.pdf.

Xiangming Gu, Soham De, Larisa Markeeva, Petar Veličković, and Razvan Pascanu. Understanding performance gap between parallel and sequential sampling in large reasoning models, 2026. URL https://arxiv.org/abs/2604.05868. arXiv preprint arXiv:2604.05868.

Daya Guo, Dejian Yang, Haowei Zhang, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025. doi: 10.1038/s41586-025-09422-z. URL https://www.nature.com/articles/s41586-025-09422-z.

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, 2026. URL https://arxiv.org/abs/2608.04001. arXiv preprint arXiv:2608.04001.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solv-

ing with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract-round2.html.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations (ICLR)*, pp. 32808–32824, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/8b4add8b0aa8749d80a34ca5d941c355-Abstract-Conference.html.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP)*, pp. 611–626, 2023. doi: 10.1145/3600006.3613165. URL https://doi.org/10.1145/3600006.3613165.

Baohao Liao, Hanze Dong, Yuhui Xu, Doyen Sahoo, Christof Monz, Junnan Li, and Caiming Xiong. Fractured chain-of-thought reasoning. In *Conference on Language Modeling (COLM)*, 2026. URL https://arxiv.org/abs/2505.12992. Accepted; arXiv:2505.12992.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison 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 (ICLR)*, pp. 39578–39601, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/aca97732e30bcf1303bc22ac3924fd16-Abstract-Conference.html.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-Refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46534–46594. Curran Associates, Inc., 2023. doi: 10.52202/075280-2019. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91edff07232fb1b55a505a9e9f6c0ff3-Abstract-Conference.html.

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. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 20275–20321. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1025. URL https://aclanthology.org/2025.emnlp-main.1025/.

Art B. Owen and Dean Eckles. Bootstrapping data arrays of arbitrary order. *The Annals of Applied Statistics*, 6(3):895–927, 2012. doi: 10.1214/12-AOAS547. URL https://doi.org/10.1214/12-AOAS547.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof Q&A benchmark. In *Conference on Language Modeling (COLM)*, 2024. URL https://arxiv.org/abs/2311.12022.

Aman Sharma and Paras Chopra. The sequential edge: Inverse-entropy voting beats parallel self-consistency at matched compute, 2025. URL https://arxiv.org/abs/2511.02309. arXiv preprint arXiv:2511.02309.

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 (ICLR)*, pp. 10131–10165, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/1b623663fd9b874366f3ce019fdfdd44-Paper-Conference.pdf.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. Challenging

BIG-Bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13003–13051. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-acl.824. URL https://aclanthology.org/2023.findings-acl.824/.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2203.11171.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems*, volume 37, pp. 95266–95290, 2024. doi: 10.52202/079017-3018. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/ad236edc564f3e3156e1b2feafb99a24-Abstract.html.

Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. Inference scaling laws: An empirical analysis of compute-optimal inference for LLM problem-solving. In *International Conference on Learning Representations (ICLR)*, pp. 55824–55845, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/8c3caae2f725c8e2a55ecd600563d172-Abstract-Conference.html.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. Technical report, Alibaba Qwen Team, 2025. URL https://arxiv.org/abs/2505.09388. arXiv preprint arXiv:2505.09388.

Shu Zhou, Rui Ling, Junan Chen, Xin Wang, Tao Fan, and Hao Wang. When more thinking hurts: Overthinking in LLM test-time compute scaling, 2026. URL https://arxiv.org/abs/2604.10739. arXiv preprint arXiv:2604.10739.

A HISTORICAL EXPERIMENTAL SETUP IN FULL

This appendix describes the historical campaign; the independent validation/test protocol is specified in Appendix M.

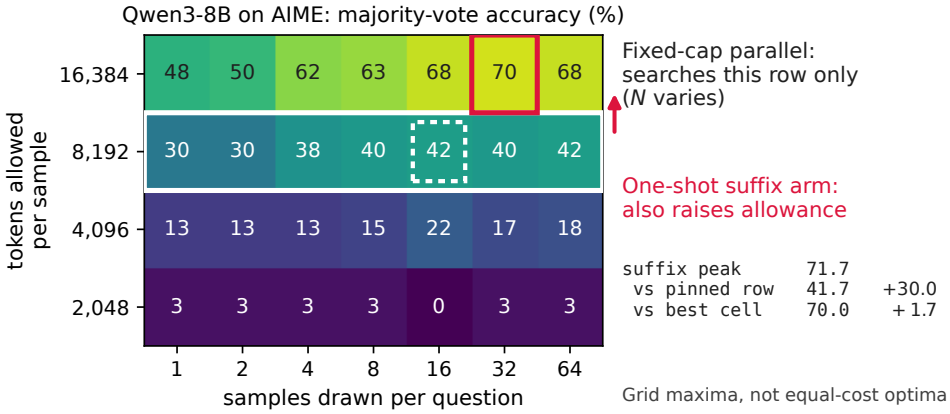

Figure 2: **The allocation choice behind a large apparent advantage.** Qwen3-8B on 60 common AIME questions. The suffix arm peaks at 71.7%; the natural-prompt row at allowance 8192 peaks at 41.7%, and the full natural-prompt grid at 70.0%. Expanding the baseline changes the gap from 30.0 to 1.7 points. These maxima measure search-space sensitivity; the independent cost-accounted comparisons are in Tables 1 and 2.

Decoding and implementation. The intended reasoning-run protocol uses temperature 0.6 and top-p 0.95 with explicit per-run allowances and sampling seeds. The source harness also has model-dependent defaults, so launch-time overrides are part of the protocol and must be retained with each run. A Qwen3-8B fallback-profile error produced 2048-token runs that were archived and excluded before re-running. The recorded environment is vLLM 0.19.1, PyTorch 2.10.0 (CUDA 12.8) and Transformers 5.6.2, on A30, L20 and RTX 5880 GPUs. Tensor parallelism varies by model and allowance, including multi-GPU 32B jobs. Hardware changes are not an independently controlled factor in these results. The historical Qwen3-8B/MATH60 wave-10 temperature remains unknown; the current ledger uses its explicit-temperature repair (Appendix L.2). This repair is not an additional prediction wave. For Qwen-14B/AIME at $c = 16384$, bank A (monolith+c2) is primary and bank B (p1–p4) is a source sensitivity check, not an independent-seed replication (Appendix L.1).

Question subsets. AIME uses 60 questions from the 2024 and 2025 sets. The original MATH-L5 suffix ladder uses a seeded random subset of 50 from a 134-question pool, whereas the allowance sweep uses a 60-question prefix. Their intersection is 22, but the current Qwen-7B and Qwen3-8B grid comparisons use returned suffix ladders on all 60 sweep questions, with the explicit-temperature repair for Qwen3-8B. Tables 3 and 5 therefore use 60 questions per cell. The sixteen historical prompt-comparison cells retain their original subsets, including 50 questions per MATH-L5 cell; the new ladders do not add independent prompt-comparison cells. Historical MATH-500 evaluations use the first 100 questions of the `HuggingFaceH4/MATH-500` test split, which packages the MATH subset released by Lightman et al. (2024). BBH also uses a 100-question subset. GPQA uses up to 198 natural-prompt questions, with suffix comparisons restricted to the available matched subset; some completed intersections contain only 31 questions. An intersection is computed once per comparison, not independently for each allowance.

Answer extraction and missingness. Open-ended correctness uses extracted answer strings and the harness's normalization/equivalence checks; multiple-choice correctness uses extracted answer labels. Voting aggregates the extracted labels, so canonicalization, equivalent answer strings and abstention handling are part of the measurement. The archived extraction audit covers the following generation banks, not every sample in the later campaign:

Task family	Extraction rate	Generations
MATH level 5	99.3%	34,304
MATH-500	99.3%	3,200
AIME	98.4%	17,408
BBH ruin-names	93.8%	3,200
GPQA Diamond	81.5%	29,056

The pooled GPQA extraction-rate difference between prompt conditions is about 0.3 points. Similar rates do not imply unbiased comparisons: missing answers can differ in correctness or failure type. Invalid answers abstain from the vote, and an all-invalid vote is incorrect. Reported intervals are conditional on this extraction pipeline.

Generation length. Historical length analyses use the character count of assembled generation text. Token allowances remain tokenizer-token limits. The Qwen3/MATH60 repair additionally records output-token IDs and termination reasons (Appendix L.2), so its generated-token counts are directly measured. Historical natural banks lack corresponding token records. Character counts are not converted to tokens or compared with the repair's token totals to infer cross-method efficiency.

Multiple-choice rendering. MATH-500 uses the same 100 questions and correct answers. Distractors are drawn from other questions' answers, preferring the same coarse type (integer, fraction, expression containing π, coordinate tuple, or other). For 95 questions, a same-type pool is available; five use a global fallback. Correct-option positions have counts 24/24/24/28. AIME is re-rendered analogously using integer answers. This construction is deterministic given its seed. It supplies candidate answers as well as a response format. We audited all 160 option sets against the original answer keys: 119 all-integer cases by exact integer comparison, and the remaining 41 MATH-500 cases by inspecting the full question and comparing the mathematical values or types of all options. No second option equivalent to the stored correct answer was found. This checks option validity relative to the benchmark keys rather than independently re-solving every benchmark question. The reproduction package includes the option sets and item-level manual notes; its script reruns the structural and integer checks.

The constructor samples answer-list positions without deduplicating values. Six MATH-500 items therefore repeat a wrong distractor, leaving three distinct candidate values despite four labels. Their IDs are `T1mc4_0005`, `0033`, `0056`, `0059`, `0073` and `0094`, with the same prefix on all six. Each retains a unique correct option. All 60 AIME option sets have four distinct integer values. Appendix G preserves the full-cohort results and reports sensitivity to these six repeated-distractor items.

B PRE-REGISTRATION DOCUMENTS

The eight recorded rounds differ in target and evidential status. Registration timestamps and queue snapshots are supporting records, not proof by themselves that all relevant data were absent. One round-1 entry followed the start of its run; round 3 uses short sequential probes before the full ladder. The appendix excerpts do not cover every original registration.

Which cohorts were predicted. The first three rounds use historical prediction cohorts, not the sixteen matched-suffix cells. They include synthetic tasks and additional model–task pairs. Feature-bank and outcome subsets can differ; for example, the wave-3 X4 record uses 96 feature questions and 50 outcome questions. A scored unit is a registered criterion, not a question, sample or gate.

Round	Wave	IDs	Cohort and counting unit
1	1	P1–P6	Six model–task predictions on AIME, MATH-L5 and GPQA.
2	2	W1–W3	1.5B/GPQA, 1.5B/MATH-L5 and 14B/BBH; three margin predictions.
3	3	X1–X6	Synthetic T5 variants and BBH across four models; six predictions, not eighteen gates.
4	4	Y1–Y4	7B and Qwen3-8B on BBH and MATH-500; four composite conditions.
5	6	Z1–Z3	Two-model MATH-500 format intervention; three conjunctions.
6	7	W1–W3	Two-model AIME format intervention; three conjunctions, distinct from wave-2 IDs.
7	8	E1–E3	7B and Qwen3-8B forcing at 8192; three historical condition slots.
8	9	F1–F3	The forcing replication at 16384; three historical condition slots.

The last two rounds use 60 AIME questions and 480 samples per model. The returned source registry contains both complete criterion blocks, reproduced in Appendix R. The source file's SHA-256 matches the returned archive manifest. This establishes the provenance of the recovered text, not independent proof of its historical blindness; the blocks contain timestamps but no frozen question-ID checksum. Wave-2 W4/W5 are unscored pending entries in the inspected record; neither is silently included in the three-outcome denominator.

Rounds 1 and 2: 4/6 and 2/3. Round 1 predicts binary crossovers from initial parallel-sample fragmentation. Four of six criteria held, with the late-entry protocol defect retained in the record. Round 2 predicts three margin bands, with two of three correct.

Round 3: 3/6 literal, 2/4 sensitivity reading. The rule adds convergence and sequential-activity gates. An omitted scope condition led the earlier analysis to remove two empirically near-saturated cells, one hit and one miss, in a separate post-hoc reading. We preserve both counts. Empirical near-saturation does not imply that a registered margin was mathematically impossible or inevitable; the sensitivity count must not be presented as a stronger prospective score.

Round 4: 4/4 literal criteria. Four registered conditions concern the suffix effect on two additional families: signs of vote changes, an ordering and range of directional discordance ratios, and shorter generated text. All four literal conditions held. A contemporaneous interpretation note warns that the sign of a near-zero estimate carries little evidence. Its claimed absence of result files conflicts with its own file count, so the note's blindness is unresolved. Only the original criteria determine the 4/4 count. The later observed separation in discordance is not a retroactive replacement for the frozen criteria.

Rounds 5 and 6: 1/3 and 0/3. The two format interventions require effects in both models, including discordance thresholds, multiplicative increases and directional ratios. All four model–task comparisons show increased discordance, but only one of the six registered criteria holds under the recorded scoring. The answer-key audit identifies a unique correct option in all items and leaves these full-cohort counts unchanged. The paired 94-question sensitivity in Appendix G is separate from the registered conjunctions.

Rounds 7 and 8: unmet gates, void downstream tests. The 8192-allowance round requires a character-length increase above $1.2\times$ in both models before interpreting its main comparative criterion. The 16384-allowance round separately requires at least one forcing per sample on average in both models, and includes a length-ratio criterion above 1.15 in both models. Its gated comparative criterion permits at most a five-point lead over the natural-grid maximum in either model. Both gates fail. The historical bookkeeping assigns zero credit to each three-condition round, yielding 14/31 overall (13/29 under the round-3 sensitivity reading). These totals combine different questions and

include void tests; they are not evidence of a 14/31 predictive classifier or six valid failures of iterative forcing. Appendix J gives the engagement measurements.

C STATISTICAL PROCEDURE

Single-cell vote comparisons use percentile question bootstraps, with 4000 resamples and a fixed seed. Task-family intervals target the pooled question–cell mean over the fixed tested cells. Within each family, question IDs are stratified by their exact observed cell-incidence pattern and sampled with replacement within strata. All model outcomes for a selected question move together, preserving shared-question dependence, cell sizes and coverage. In table order, there are 100, 50, 60, 50 and 100 unique questions. AIME has incidence strata of 32 and 28 questions; GPQA has 31 and 19; the other families each have one stratum. No stratum is a singleton. Each family resets the generator to seed 20260816; incidence tuples and question IDs are sorted canonically. Endpoints are draws 100 and 3899 (zero-based) of 4000 sorted replicates. These nominal intervals condition on the models, incidence patterns, recorded banks and extraction; they do not propagate generation-seed variability or adjust for multiple comparisons. This shared-question procedure replaces the earlier cell-then-question resampling, with all point estimates and comparison cohorts unchanged.

Sample-index pairing aligns separately generated continuations for the discordance diagnostic. It does not hold the random continuation fixed or identify a within-trajectory intervention. For Table 6, the vote change averages paired question-level differences pooled across the family's cells, so cells with more shared questions carry greater weight. The table's n_q counts question–cell observations, not unique questions across models. D_0 pairs natural sample IDs 0–7 with 8–15, whereas D_1 pairs natural IDs 0–7 with suffix IDs 0–7. Each discordance pools discordant pair counts over total matched pairs; $E = D_1 - D_0$ is formed on the same question–cell observations. Appendix L.3 reports the per-cell controls and their question-bootstrap intervals. In the auxiliary Table 9, R divides the pooled natural-wrong/suffix-correct count by its opposite count. The character-length change instead equally averages the cells' relative changes in mean sample length; it is not a ratio of pooled lengths.

All five cells in Table 5 use 60 matched questions. Its 95% bootstrap ranges condition on the selected configurations and omit argmax-selection uncertainty; they are not nominal confidence intervals for the configuration-selection procedure. A separate 400-split half-sample selection analysis chooses a configuration on one half and scores it on the other, then swaps halves. Its variation across splits is not a population confidence interval, because the same questions recur. Mean suffix-minus-grid margins are $+3.78$, $+6.20$ and -6.85 points for Qwen3-8B, Qwen-14B (bank A) and Qwen-7B on AIME, respectively; bank B gives $+6.18$ for Qwen-14B. On MATH-L5 the corresponding means are -0.47 for the repaired Qwen3-8B cell and $+0.39$ for Qwen-7B. The repaired cell's mean suffix-minus-row margin is $+9.37$ points. These are selection-sensitivity summaries, not proof that selection bias is eliminated or that the strategies are equivalent.

A corrected reproduction must use one question intersection per comparison, validate complete sample identifiers and allowances, and resolve duplicate shards explicitly. Script execution alone does not validate those invariants.

D CORRECTIONS TO OUR EARLIER VERSION

Length and termination. An earlier truncation proxy compared character length with a token allowance. Character-to-token ratios vary too much for that threshold to identify termination. An explicit end-of-thinking check gave a natural-arm completion rate of 83.9%, rather than the earlier proxy's 53.6%. The old truncation claim is withdrawn; a reasoning-close marker is itself distinct from final generation termination.

A predictive signal. The fragmentation-based selector is not validated by the prospective records. Appendix Q reports the failed rules without advertising a zero-cost allocation method or a universal impossibility result.

Baseline ledger. An earlier implementation compared maxima on different question sets. In particular, a 32-question Qwen-14B/AIME grid result was subtracted from a 60-question suffix result,

and MATH-L5 mixed a 50-question suffix subset with a 22-question cross-arm intersection. Using common questions changes the Qwen-14B grid-relative margin from -7.9 to $+3.3$ points. The earlier claim that three of four comparisons reverse is withdrawn. Table 5 now uses all 60 matched questions in each of five cells, including the returned Qwen-7B/MATH-L5 ladder and the Qwen3-8B/MATH-L5 explicit-temperature repair. The latter's suffix, fixed-row and full-grid peaks are 76.7%, 68.3% and 78.3%, giving margins $+8.3$ and -1.7 points; the Qwen-7B/MATH-L5 peaks all equal 73.3%. The five grid-relative signs are two negative, one zero and two positive, not a count of significant reversals.

Attribution and diagnostic definitions. The Qwen3-8B/AIME gap changes from 30.0 to 1.7 points when the parallel search space expands, but those maxima are not cost-matched and their difference is not a causal decomposition. The earlier label *rewrite volume* is replaced by correctness discordance. The historical D_1 estimates are retained, but the natural–natural D_0 control shows substantial no-suffix discordance and a much smaller excess E (Appendix L.3). Historical length changes remain character measurements; the repair's token logs are a separate exception (Appendix L.2).

E Cell ledger and paired correctness statistics

The grid-comparison rows use the current 60-question matched banks, including the repaired Qwen3-8B/MATH-L5 suffix. Other rows retain historical comparison subsets. Qwen-7B/MATH-L5 also retains its original 50-question row for traceability; it is distinct from the current 60-question row. A missing grid comparison is not zero, and regime labels from different subsets are not interchangeable. Table 8 preserves all sixteen historical prompt-comparison cells, including their original 50-question MATH-L5 subsets.

Table 7: Recorded maxima, in percent and percentage-point differences. Row and grid maxima are descriptive, not cost-matched. Qwen-7B/MATH-L5 has two rows for distinct historical (50) and current (60) question subsets. acc@1 is the first natural sample's accuracy at $c = 8192$. Missing grid entries denote unavailable same-subset comparisons.

Task	Model	n_q	acc@1	Row peak	Suffix peak	vs row	vs grid
AIME	Qwen3-8B	60	30.0	41.7	71.7	+30.0	+1.7
AIME	Qwen-14B	60	36.7	58.3	73.3	+15.0	+3.3
AIME	Llama-8B	32	25.0	53.1	62.5	+9.4	n/a
AIME	Qwen-7B	60	35.0	46.7	55.0	+8.3	−6.7
AIME	Qwen-1.5B	60	18.3	33.3	40.0	+6.7	n/a
MATH-L5	Qwen3-8B	60	61.7	68.3	76.7	+8.3	−1.7
MATH-L5	Qwen-14B	50	58.0	70.0	74.0	+4.0	n/a
MATH-L5	Llama-8B	50	54.0	70.0	70.0	0.0	n/a
MATH-L5	Qwen-7B	50	70.0	70.0	66.0	−4.0	n/a
MATH-L5	Qwen-7B	60	66.7	73.3	73.3	0.0	0.0
MATH-500	Qwen-7B	100	66.0	71.0	73.0	+2.0	n/a
MATH-500	Qwen3-8B	100	76.0	76.0	75.0	−1.0	n/a
GPQA	Qwen-14B	31	51.6	58.1	58.1	0.0	n/a
GPQA	Llama-8B	31	38.7	58.1	58.1	0.0	n/a
GPQA	Qwen-7B	50	42.0	60.0	56.0	−4.0	n/a
BBH-ruin	Qwen3-8B	100	66.0	72.0	73.0	+1.0	n/a
BBH-ruin	Qwen-7B	100	40.0	56.0	47.0	−9.0	n/a

F Effect on aggregation

On GPQA the pooled vote change is -8.0 points, compared with -1.7 for individual-sample accuracy. On one BBH cell the benefit of voting over an individual sample changes from 11.6 to 9.1 points; on the other it changes from 10.6 to 10.7. Thus the adverse aggregation pattern is not uniform. These are descriptive comparisons at different aggregation levels. Near-symmetric correctness discordance

Table 8: Historical natural–suffix discordance D_1 and directional count ratio R for all sixteen matched cells at $c = 8192$, $N = 8$. The raw D_1 ranges do not overlap between these task groups; the ratios do. Stars mark open-ended ratios inside the multiple-choice range. These are separate-run pairs, not within-trajectory revisions. Per-cell natural–natural D_0 and excess E are reported in Appendix L.3.

Answer space	Cell	n_q	D_1	R
open-ended	AIME / Llama-8B	32	24.2%	1.58
open-ended	AIME / Qwen-7B	60	18.5%	0.89*
open-ended	AIME / Qwen-14B	60	18.8%	1.57
open-ended	AIME / Qwen-1.5B	60	15.0%	1.57
open-ended	AIME / Qwen3-8B	60	13.5%	2.10
open-ended	MATH-L5 / Llama-8B	50	15.0%	0.67
open-ended	MATH-L5 / Qwen-14B	50	14.0%	2.11
open-ended	MATH-L5 / Qwen-7B	50	11.2%	0.80
open-ended	MATH-L5 / Qwen3-8B	50	9.5%	2.80
open-ended	MATH-500 / Qwen-7B	100	7.9%	1.86
open-ended	MATH-500 / Qwen3-8B	100	5.0%	1.86
4-way choice	BBH-ruin / Qwen-7B	100	38.0%	0.92
4-way choice	GPQA / Qwen-14B	31	34.3%	0.93
4-way choice	GPQA / Qwen-7B	50	33.8%	0.90
4-way choice	GPQA / Llama-8B	31	33.5%	0.89
4-way choice	BBH-ruin / Qwen3-8B	100	31.1%	1.06

Table 9: Auxiliary historical prompt diagnostics for the same sixteen cells as Table 6. R pools directional pair counts; character changes equally average cell-wise relative mean-length changes. Neither quantity measures token efficiency or a within-trajectory revision.

Task family	R	Δ chars
MATH-500	1.86	-11.4%
MATH level 5	1.24	-13.3%
AIME	1.44	-3.0%
GPQA Diamond	0.91	-18.7%
BBH ruin-names	0.98	-40.0%

alone does not establish that wrong answers become more dispersed or explain the vote change; that mechanism would require the full answer-frequency distributions.

G THE FORMAT INTERVENTION

The full-cohort results use 100 MATH-500 and 60 AIME questions per model, including all six repeated-wrong-distractor items. The observed increase in discordance is distinct from the six registered threshold tests, only one of which succeeds.

Distinct-option sensitivity. We exclude the six MATH-500 items identified by option duplication, together with their paired open-ended versions, from both models. Selection uses the option sets, not model outcomes. Table 11 shows that the increase persists on the remaining 94 questions. This is a post-hoc sensitivity analysis; it replaces neither the original table nor the registered denominators. AIME retains all 60 questions and its original values because no distractors repeat. Answer cues, extraction and effective difficulty still change between renderings.

Table 10: Correctness discordance under open-ended and multiple-choice renderings of the same questions. The factors are descriptive; adding answer options also changes cues, extraction and potentially difficulty. Baseline accuracy is measured without the suffix at $N = 8$.

Questions	Model	D, open	D, options	Factor	Baseline accuracy
MATH-500	Qwen-7B	7.9%	47.5%	6.0×	$70.0 \to 71.0$
AIME	Qwen-7B	18.5%	27.1%	1.5×	$43.3 \to 40.0$
MATH-500	Qwen3-8B	5.0%	23.8%	4.8×	$76.0 \to \mathbf{94.0}$
AIME	Qwen3-8B	13.5%	22.1%	1.6×	$40.0 \to \mathbf{63.3}$

Table 11: Paired MATH-500 sensitivity on items with four distinct option strings. Correctness discordance D_1 uses eight samples per question and prompt. The same 94 questions enter both renderings and both models.

Model	n_q	D_1, open	D_1, options	Factor
Qwen-7B	94	7.98%	46.81%	5.87×
Qwen3-8B	94	5.32%	23.94%	4.50×

H Empirical modes and finite-sample calibration

Finite estimation banks can misidentify the population mode. Resolving an empirical tie by first occurrence does not establish a unique population mode, and vote@32 is not a finite-sample upper bound. The asymptotic argument requires independent fixed-distribution draws and an explicit treatment of population modal ties.

Predictors and error metric. At each question and allowance, samples 0–31 estimate answer frequencies; 400 Monte Carlo votes at each $N \in \{1, 2, 4, 8, 16, 32\}$ predict voting accuracy, using the historical random seed 20260816 and iteration order. Observed votes use prefixes of samples 32–63. The first-sample constant uses one estimation draw; the 32-sample constant averages correctness over all 32, including invalid answers as incorrect. Both repeat that prediction across N. Replay instead votes on the first N estimation draws. All voting uses abstention for invalid extractions and first-occurrence tie resolution. For each of 144 cell-allowance-count points, predicted and observed accuracies are first averaged over its 60 questions; their absolute difference is then averaged equally across points. This is not the mean of per-question absolute errors.

Paired question uncertainty. This post-hoc sensitivity analysis retains the tested cells, recorded sample banks and Monte Carlo predictions. Each bank has 120 unique questions: 60 AIME questions shared across four models and 60 MATH-L5 questions shared across two. We draw 20000 stratified question bootstraps with NumPy PCG64 seed 20260915. A question's multiplicity moves jointly across every model, allowance, count and predictor; task and cell-incidence strata preserve the question count at every point. The paired error difference is baseline MAE minus model MAE in each draw, not a subtraction of marginal interval endpoints. Banks A and B are provenance-defined alternatives (Appendix L.1), not independent replications.

Table 13 reports nominal pointwise 95% percentile intervals and basic reflected intervals, computed as $[2\widehat{\Delta} - q_{.975},\ 2\widehat{\Delta} - q_{.025}]$ from the same draws. The latter were added after inspecting the initial percentile analysis. Absolute aggregate error is non-smooth near zero, so resampled errors can be shifted relative to the observed error. Reflection is a sensitivity check, not a guaranteed correction. Both forms are reported for all contrasts without multiplicity adjustment; neither includes fresh-generation, selection or Monte Carlo-prediction uncertainty. The improvement over the 32-sample constant persists under both interval forms and banks. The smaller replay difference is sensitive to interval construction, supporting descriptive calibration rather than a claim of established superiority to replay.

Table 12: Empirical modal-answer accuracy from the first 32 natural samples (IDs 0–31), in percent, on 60 questions per cell. Invalid extractions abstain, all-invalid votes are incorrect, and tied modes use first occurrence, so these estimates equal vote@32. Qwen-14B uses bank A; bank B gives the same rounded row. Held-out finite-N validation uses IDs 32–63 separately.

Model	Task	2048	4096	8192	16384
Qwen-7B	AIME	11.7	31.7	45.0	60.0
Qwen-14B	AIME	11.7	26.7	53.3	70.0
Qwen3-8B	AIME	3.3	16.7	40.0	70.0
Qwen-32B	AIME	10.0	31.7	55.0	71.7
Qwen3-8B	MATH level 5	28.3	48.3	68.3	78.3
Qwen-7B	MATH level 5	43.3	58.3	**73.3**	**73.3**

Table 13: Post-hoc calibration sensitivity for both provenance-defined 14B banks. Positive paired MAE differences favor the empirical-distribution model. Both nominal 95% question-bootstrap interval forms are shown; the basic interval was added after inspecting the initial percentile analysis. The banks are alternatives, not independent replications.

14B bank	Model	First-sample constant	32-sample constant	Replay
A: main	2.175	6.435	6.260	2.616
B: alternative	2.130	6.319	6.076	2.639

Bank	Baseline minus model	Difference (pp)	Percentile 95%	Basic 95%
A	Constant: first sample	+4.261	[+2.97, +5.43]	[+3.09, +5.56]
A	Constant: 32-sample mean	+4.086	[+2.87, +4.48]	[+3.69, +5.30]
A	Replay vote@N	+0.441	[+0.17, +1.52]	[-0.64, +0.72]
B	Constant: first sample	+4.189	[+2.90, +5.41]	[+2.97, +5.48]
B	Constant: 32-sample mean	+3.946	[+2.72, +4.36]	[+3.53, +5.17]
B	Replay vote@N	+0.509	[+0.21, +1.56]	[-0.54, +0.80]

I SENSITIVITY TO EQUAL-MULTIPLIER ANCHORS

Starting from $(N, c) = (1, c)$, compare $(1, kc)$ with (k, c). The two have equal nominal token ceilings kc, not necessarily equal realized token use. Each of the six natural-grid cells uses 60 questions shared across its compared configurations. The table is recalculated from the returned accuracy vectors using bank A for Qwen-14B.

Start	k	Cells	Ratio for positive sample gain	Sample gain ≤ 0	Both zero
2048	2	6	$16.0\times$	5	0
2048	4	6	9.5–$15.0\times$	4	0
2048	8	6	5.0–$10.5\times$	3	0
4096	2	6	$7.0\times$	5	0
4096	4	6	4.25–$7.67\times$	2	0
8192	2	6	$5.0\times$	4	1

The penultimate column counts cases where allowance gains are positive and sample-count gains are zero or negative; it must not be labeled as an exactly-zero gain. Allowance gains are larger in 35/36 comparisons under either bank, with both zero for Qwen-7B/MATH-L5 at $c = 8192$, $k = 2$. Bank B leaves all counts and other ratio ranges unchanged, but gives 4.25–$6.67\times$ at $c = 4096$, $k = 4$ (Appendix L.1). These are overlapping comparisons, not independent replications. Starting at one sample and resolving tied votes by first occurrence can make a two-sample vote identical to the first sample when both are valid. The check therefore does not establish anchor-independent or uniformly cost-optimal depth dominance.

J TWO ATTEMPTS AT AN ITERATIVE-FORCING CONDITION

We implemented generation followed by suppression of the reasoning-close marker, a continuation phrase, and further generation, up to the allowance or four forcing events. This differs both from our one-shot suffix and from multi-attempt refinement with accumulated context in other work.

Allowance	Model	Forcings/sample	Never forced	Character-length ratio
8192	Qwen-7B	0.39	291/480	$0.99\times$
8192	Qwen3-8B	0.27	350/480	$1.01\times$
16384	Qwen-7B	0.71	161/480	$0.96\times$
16384	Qwen3-8B	0.59	198/480	$1.02\times$

The 8192 round fails its $1.2\times$ length gate; the 16384 round fails its mean-forcing gate of at least one, and does not meet its 1.15 length ratio. Forcing nevertheless occurs in some samples. At 16384, most samples receive at least one forcing event, so the failure is not equivalent to no forcing or an implementation that never engages. These observations do not establish that larger budgets are necessary or that another implementation cannot work.

For completeness, the 16384 vote accuracies are 51.7 and 65.0 for forcing, versus 55.0 and 71.7 for the corresponding one-shot suffix. Natural-grid maxima are 61.7 and 70.0, but are not a cost-matched forcing baseline. These descriptive numbers do not reinstate the void registered efficacy tests. The smaller-allowance and larger-allowance rounds are retained as separate attempts.

K SUFFIX WORDING

For Qwen-7B/AIME, the main suffix and two alternatives (*Let me double-check my reasoning so far.* and *Hmm, I should verify this step by step.*) give correctness-discordance rates 18.5%, 18.8% and 16.2%, with directional ratios 0.89, 0.88 and 1.29. This single-cell sensitivity check suggests a more stable discordance rate than directional balance under these wordings; it does not establish generality over prompts.

L VALIDATION OF SAMPLING BANKS AND MATCHED CONTROLS

L.1 EXECUTION-SOURCE BANKS FOR THE 14B GRID

The Qwen-14B/AIME natural condition at $c = 16384$ has two complete execution-source banks, defined from provenance before accuracy calculation. Bank A combines the original monolith's first 32 questions with the remaining 28 questions in c2; bank B combines p1–p4, each covering 15 questions. Each contains $60 \times 64 = 3840$ unique question–sample records, with no within-bank overlap. The main text uses bank A (monolith+c2); bank B is the sensitivity alternative. The banks share the nominal seed and differ in execution source and batching. Other allowances use the same sources in both analyses.

At $N = 64$, extending $c = 8192$ to 16384 increases accuracy from 58.33% to 68.33% in A and 66.67% in B, giving extra-depth gains of $+10.00$ and $+8.33$ percentage points (pp). Across $c \in \{2048, 4096, 8192, 16384\}$ and $N \in \{1, 2, 4, 8, 16, 32, 64\}$, both natural-grid peaks are 70.00%; the lowest-N maximizing settings are $(c, N) = (16384, 16)$ for A and $(16384, 32)$ for B. The suffix peak is 73.33%, a $+3.33$ pp gap under either bank.

For distribution-model calibration, only the 14B execution-source bank changes between the two analyses. Table 13 reports the empirical-distribution, first-sample constant, 32-sample constant and replay errors, together with paired question-bootstrap sensitivity ranges. The estimation/evaluation split and dependence structure are specified in Appendix H.

L.2 EXPLICIT-TEMPERATURE REPAIR ON QWEN3/MATH60

The historical Qwen3-8B/MATH-L5 suffix run has actual temperature `UNKNOWN`. Its repair uses explicit temperature 0.6, top-$p = 0.95$, sampling seed 42 and $N = 8$ at all seven al-

Table 14: Repaired Qwen3/MATH60 results and measured output cost. Each row contains 480 samples; all seven rows total 12,824,834 generated tokens, counted from saved output-token IDs. `stop` and `length` are termination reasons, not correctness labels. Accuracy maxima are selected on the same 60 questions. Token accounting covers only the repaired suffix arm: historical natural character counts are neither converted to nor compared with these token counts, so the accuracy comparisons are not equal-realized-cost comparisons.

c	Correct	Generated tokens	`stop`	`length`
512	7/60	245,760	0	480
1024	12/60	490,994	2	478
2048	16/60	961,368	56	424
4096	30/60	1,727,919	153	327
8192	43/60	2,714,709	327	153
12288	43/60	3,172,022	397	83
16384	46/60	3,512,062	423	57

lowances in Table 14, matching the logged natural sampling settings. It retains the same 60 questions, their order and the original suffix. The repaired model revision is fixed to `b968826d9c46dd6066d109eabc6255188de91218`; the historical revision was unpinned and its match is unknown. This is a configuration repair, not an additional prediction wave or an estimate of the causal temperature effect. Voting retains the frozen mathematical-equivalence labels, discards invalid extractions and resolves ties by first occurrence.

The repaired suffix reaches $46/60 = 76.67\%$ at $(c, N) = (16384, 8)$, matching natural prompting at the same N and cap. The correct-question sets differ by one exclusive success in each arm. The natural fixed-$c = 8192$ row peaks at $41/60 = 68.33\%$ with $N = 32$; the full natural grid peaks at $47/60 = 78.33\%$ with $(c, N) = (16384, 16)$. The suffix advantage is therefore $+8.33$ pp against the fixed row and -1.67 pp against the full grid.

L.3 Natural–natural correctness-discordance control

At $c = 8192$, let $u_{q,s}$ and $v_{q,s}$ be binary correctness labels for natural and suffix outputs. Natural group A uses sample IDs 0–7 and group B uses IDs 8–15; suffix IDs 0–7 are paired with A. These group names are unrelated to the execution-source banks above. For a cell with n_q questions, define

$$D_0 = \frac{1}{8n_q} \sum_{q=1}^{n_q} \sum_{s=0}^{7} \mathbf{1}\{u_{q,s} \neq u_{q,s+8}\},$$

$$D_1 = \frac{1}{8n_q} \sum_{q=1}^{n_q} \sum_{s=0}^{7} \mathbf{1}\{u_{q,s} \neq v_{q,s}\}, \qquad E = D_1 - D_0.$$

Sample-index pairing defines a correctness diagnostic rather than holding a random continuation fixed. We retain the historical 16 natural/suffix source intersections without outcome filtering: Llama-8B/AIME has 32 questions; GPQA has 31 for Llama-8B and Qwen-14B, and 50 for Qwen-7B; the four historical MATH-L5 cells each retain 50. The newer 60-question cohorts, including the repaired run, are not appended as independent cells or substituted for these denominators.

What excess discordance identifies. For one question, suppose independent natural and suffix draws have correctness probabilities p_n and p_s. Expected natural/suffix discordance is $p_n(1 - p_s) + (1 - p_n)p_s$; expected natural/natural discordance is $2p_n(1 - p_n)$. Thus the expected per-question excess, in probability units, is

$$\mathbb{E}[E_q] = (1 - 2p_n)(p_s - p_n).$$

It can vanish when $p_n = 1/2$ despite a change in correctness probability, and contributions can cancel across questions. The control calibrates correctness discordance; a zero-crossing excess interval is not evidence of equivalent prompt effects. The separately reported vote-accuracy contrast addresses the accuracy effect. Neither discordance statistic counts changes between two incorrect answers.

Table 15: Historical 16-cell natural–natural control. D_0 and D_1 are percentages; $E = D_1 - D_0$ and its 95% confidence interval (CI) are in pp. Intervals use 4000 paired question-bootstrap resamples of the within-question excess, conditional on recorded outputs, sample-index pairing, extraction and scoring; they are cellwise and uncorrected for multiple comparisons. n_q is the retained source-intersection denominator. Qwen-* and Llama-8B abbreviate DeepSeek-R1 distilled models; Qwen3-8B denotes the Qwen3 model.

Task	Model	n_q	D_0	D_1	E	Conditional 95% CI
MATH-500	Qwen-7B	100	8.50	7.88	-0.63	$[-3.00, +1.50]$
MATH-500	Qwen3-8B	100	5.75	5.00	-0.75	$[-2.38, +0.75]$
MATH-L5	Qwen-7B	50	9.25	11.25	$+2.00$	$[-1.50, +6.00]$
MATH-L5	Qwen-14B	50	13.00	14.00	$+1.00$	$[-1.50, +4.00]$
MATH-L5	Qwen3-8B	50	9.00	9.50	$+0.50$	$[-2.25, +3.50]$
MATH-L5	Llama-8B	50	16.00	15.00	-1.00	$[-5.00, +2.75]$
AIME	Qwen-1.5B	60	13.33	15.00	$+1.67$	$[-2.29, +5.83]$
AIME	Qwen-7B	60	14.58	18.54	$+3.96$	$[+0.21, +8.13]$
AIME	Qwen-14B	60	17.71	18.75	$+1.04$	$[-2.08, +4.58]$
AIME	Qwen3-8B	60	13.75	13.54	-0.21	$[-3.54, +2.92]$
AIME	Llama-8B	32	22.27	24.22	$+1.95$	$[-4.30, +8.59]$
GPQA	Qwen-7B	50	34.75	33.75	-1.00	$[-6.75, +5.25]$
GPQA	Qwen-14B	31	28.23	34.27	$+6.05$	$[-0.40, +12.90]$
GPQA	Llama-8B	31	33.06	33.47	$+0.40$	$[-6.05, +6.85]$
BBH	Qwen-7B	100	35.13	38.00	$+2.88$	$[-1.38, +7.38]$
BBH	Qwen3-8B	100	29.13	31.13	$+2.00$	$[-1.63, +5.63]$

M INDEPENDENT CONFIRMATION: PROTOCOL AND COMPLETE RESULTS

M.1 DATA, SELECTION AND EXECUTION

We use the test split of the `EleutherAI/hendrycks_math` distribution at revision `21a5633873b6a120296cce3e2df9d5550074f4a3`. Among 5000 source items, 1324 are level 5; excluding canonical exact matches to historical task manifests leaves 1190. A frozen salted-hash selection chooses 600 without generation outcomes, splitting 200 validation and 400 held-out questions. This split is new to the audited campaign manifests. The 400 held-out questions are not used for selecting configurations or rewriting prompts. Validation seeds are 1101, 1102, 1103; test seeds are 2101, 2102, 2103. Model-specific selection receipts precede their test generation and reconcile exactly with the complete two-model selector.

Natural sampling evaluates $N \in \{1, 2, 4, 8, 16, 32\}$ and $c \in \{2048, 4096, 8192, 16384\}$. Suffix sampling evaluates the same four allowances at $N = 8$. Self-Refine evaluates solution allowances 4096 and 8192, trajectory counts $\{1, 2, 4, 8\}$, and one or two critique/revision rounds, with a 2048-token feedback allowance. This yields 44 distinct configurations per model; pinned and depth select subsets of the natural grid. Selection maximizes validation mean accuracy within each model, strategy and target, then minimizes mean cost and configuration ID to break ties. All-invalid votes remain incorrect, and plurality ties use first occurrence. The same frozen extraction and equivalence rule is used for all strategies, with numeric absolute/relative tolerance 10^{-3}, a symbolic-expression allow-list, and a ten-second CPU grading timeout counted as incorrect.

A concrete iterative comparator. Our adapted Self-Refine uses full previous outputs in a new user message, retaining earlier reasoning across native chat templates. Its feedback instruction is:

> Review the current solution step by step. Check the mathematics, assumptions and missing cases. Identify specific errors and give actionable corrections if you find them. If the reasoning is sound, explain what you checked instead of inventing an error. Give mathematical feedback, not a confidence score.

The revision instruction is:

> Use the previous solution and feedback to solve the original problem again. Keep justified steps, correct any errors, and independently verify the conclusion. Give a complete solution and put the final answer inside `\boxed{}`.

The task-specific prompts are frozen adaptations of the iterative call graph, not the original paper's few-shot prompts. Both models select a single trajectory and one round at solution allowance 8192 for the primary target. The suffix intervention instead appends the same historical reconsideration string after the rendered prompt and does not consume earlier solutions.

Model and runtime identity. Qwen3-8B uses revision `b968826d9c46dd6066d109eabc6255188de91218`; the R1-distilled Qwen-7B uses `916b56a44061fd5cd7d6a8fb632557ed4f724f60`. Both use temperature 0.6, top-p 0.95, native end-of-sequence tokens, bfloat16, single-GPU execution, and a 40960-token context. The frozen runtime uses vLLM 0.19.1 (Kwon et al., 2023), PyTorch 2.10.0 and Transformers 5.6.2, eager execution, disabled prefix caching and chunked prefill. Logical requests have deterministic hash-derived seeds incorporating model, phase, replicate, question, method, allowance, sample and stage. Hardware smoke tests precede mixed execution on L20, RTX 5880, A30 and RTX 3090; model weights, decoding and scientific settings are held fixed. Hardware allocation follows locked question-shard plans, not observed accuracy.

Costs and returned evidence. Validation comprises 288000 logical calls in 48 units; the selected test union comprises 110400 calls in 96 units. Shared generated samples are reused across identical selected policies, but each policy is charged its own full input and output tokens, including repeated critique/revision context. Unique-generation cost, per-policy logical cost, and allocated GPU time are separate ledgers. Allocated time is an execution expense, not an interactive latency measurement. All 34800 available policy scores remain in the analysis.

Table 16: Every model/seed primary accuracy (%). Each row uses the same 400 questions. q3 denotes Qwen3-8B and q7 denotes DeepSeek-R1-Distill-Qwen-7B.

Model	Seed	Joint	Pinned	Depth	Suffix	Self-Refine
q3	2101	75.75	66.25	74.25	53.00	73.00
q3	2102	75.75	65.00	74.25	52.75	76.00
q3	2103	76.50	63.25	72.50	52.25	74.00
q7	2101	72.00	68.25	69.75	59.25	67.00
q7	2102	71.75	68.25	69.25	60.25	66.00
q7	2103	72.25	67.50	69.25	60.50	68.00

Table 17: Every model/seed primary mean actual token cost per question, including all logical input and output calls. Rows correspond to Table 16.

Model	Seed	Joint	Pinned	Depth	Suffix	Self-Refine
q3	2101	30,062	24,065	7,523	28,912	21,347
q3	2102	30,249	24,068	7,441	29,077	21,261
q3	2103	30,046	23,978	7,622	29,017	21,298
q7	2101	21,074	17,758	5,228	25,379	17,737
q7	2102	21,668	17,604	5,392	25,535	17,745
q7	2103	21,218	17,569	5,225	25,550	17,632

Complete slim records preserve actual token counts and dependency hashes. The locally returned raw subset contains sample zero for the lexicographically smallest question ID in each unit, across all generated banks and stages. This deterministic export rule differs from the originally described source-order first question in 90 of 96 units; it does not select on outcomes or alter the complete score grid. Full raw outputs remain inventoried by hash; neither those outputs nor the raw subset is included in the anonymous numerical package. Independent local reduction reproduces the logical costs and primary bootstrap statistics; full symbolic grading uses the authenticated Linux scores rather than a second local implementation.

M.2 Registered comparisons and uncertainty

We average equally over the two fixed models, three seeds and 400 questions. There are 400 question clusters, not 2400 independent questions. For each joint-minus-baseline contrast, the cost ratio is the ratio of pooled mean costs. The registered practical-superiority criterion requires an accuracy-difference lower endpoint above zero and a cost-ratio upper endpoint at most 1.05. At the primary target, all four accuracy intervals are positive, but only the suffix comparison meets the cost requirement.

We use 20000 percentile-bootstrap draws with seed 20260912 and NumPy PCG64 streams for question and seed indices. Paired question resampling preserves all models and strategies; crossed resampling additionally draws three seed indices with replacement, jointly across models and strategies. This is motivated by the crossed-factor bootstrap of Owen & Eckles (2012). The reported interval uses the smaller lower and larger upper endpoints of these two procedures. Each of eight marginal intervals has nominal confidence 99.375%, using Bonferroni adjustment for a 95% family level. These are bootstrap approximations with only three seeds, conditional on validation-selected policies; they do not include uncertainty from rerunning validation selection.

Table 1 in the main text reports the complete primary interval set.

Joint exceeds the suffix in accuracy in all six model/seed cells. It does not exceed Self-Refine in every cell: Qwen3 seed 2102 scores 75.75% for joint and 76.00% for Self-Refine. Leaving out any one seed leaves every pooled primary accuracy difference positive, while the three cost disadvantages remain.

Table 18: Complete descriptive results at both secondary selection targets. The column target constrains validation mean cost; tokens/q. is actual held-out input-plus-output cost. Means use both models, three seeds and 400 questions. The low-target suffix aggregate remains NA because its Qwen3 slot is infeasible.

Strategy	16384: Acc. (%)	Tokens/q.	65536: Acc. (%)	Tokens/q.
Joint	71.67	10,185	74.46	36,369
Pinned	62.33	8,218	68.00	41,773
Single long solve	71.54	6,405	71.54	6,405
One-shot suffix	NA	NA	75.29	45,718
Adapted Self-Refine	67.42	15,141	71.67	50,140

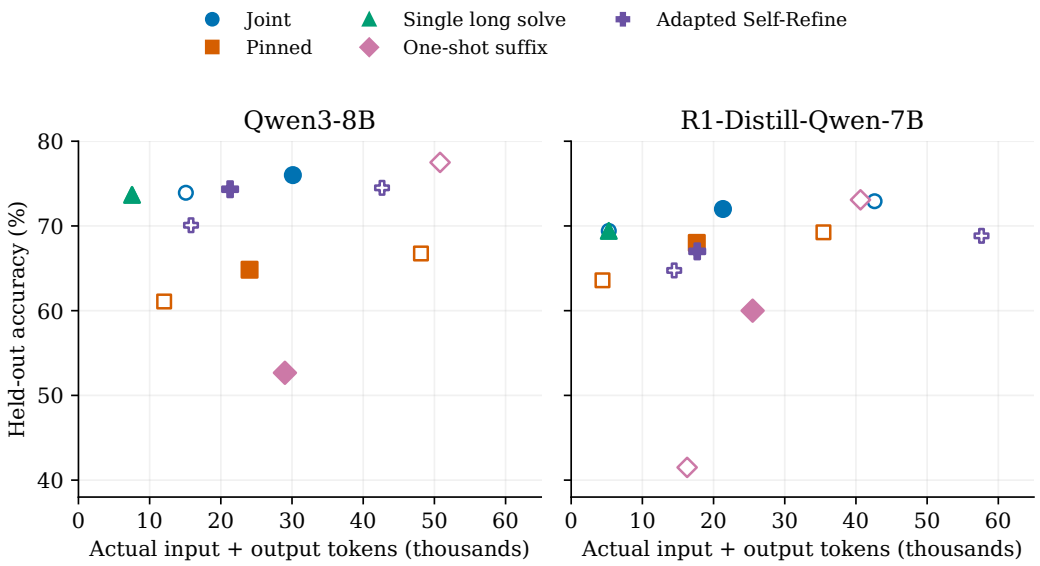

Figure 3: Held-out accuracy and actual input-plus-output costs, averaged over 400 questions and three test seeds per model. Filled points mark the primary selection target 32768; hollow points mark 16384 and 65536. Identical selected policies can overlap; Qwen3's infeasible low-target suffix is absent. Table 19 gives complete target means.

M.3 All targets and selected settings

The 16384 and 65536 targets remain descriptive secondary analyses. Qwen3's suffix has no feasible configuration at 16384, so its model result and any two-model aggregate requiring it are NA. Qwen-7B's suffix remains available at that target; its result is not substituted for the missing model. The missing slot accounts for 1200 policy scores, giving 34800 rather than 36000. A CPU analysis amendment made this pre-existing selector status explicit without changing the primary sample, contrasts or interval rule.

Figure 3 plots the selected settings separately by model. It does not interpolate a frontier or reselect using test data.

Table 19: All three targets, equally pooled over models and seeds. The 16384 and 65536 targets are descriptive; the primary target remains 32768. The low-target suffix aggregate is NA because its Qwen3 validation slot is infeasible.

Selection target	Strategy	Accuracy (%)	Tokens/question
16384	Joint	71.67	10,185
16384	Pinned	62.33	8,218
16384	Single long solve	71.54	6,405
16384	One-shot suffix	NA	NA
16384	Adapted Self-Refine	67.42	15,141
32768	Joint	74.00	25,720
32768	Pinned	66.42	20,840
32768	Single long solve	71.54	6,405
32768	One-shot suffix	56.33	27,245
32768	Adapted Self-Refine	70.67	19,503
65536	Joint	74.46	36,369
65536	Pinned	68.00	41,773
65536	Single long solve	71.54	6,405
65536	One-shot suffix	75.29	45,718
65536	Adapted Self-Refine	71.67	50,140

Table 20: All 30 frozen selection slots. Zero rounds denotes a one-shot policy; one or two denotes the number of feedback/revision rounds. The single infeasible slot remains explicit.

Model	Target	Strategy	c	N	Rounds
q3	16384	Joint	16384	2	0
q3	16384	Pinned	8192	2	0
q3	16384	Single long solve	16384	1	0
q3	16384	One-shot suffix	NA	NA	NA
q3	16384	Adapted Self-Refine	4096	1	1
q3	32768	Joint	16384	4	0
q3	32768	Pinned	8192	4	0
q3	32768	Single long solve	16384	1	0
q3	32768	One-shot suffix	4096	8	0
q3	32768	Adapted Self-Refine	8192	1	1
q3	65536	Joint	16384	4	0
q3	65536	Pinned	8192	8	0
q3	65536	Single long solve	16384	1	0
q3	65536	One-shot suffix	16384	8	0
q3	65536	Adapted Self-Refine	8192	2	1
q7	16384	Joint	16384	1	0
q7	16384	Pinned	8192	1	0
q7	16384	Single long solve	16384	1	0
q7	16384	One-shot suffix	2048	8	0
q7	16384	Adapted Self-Refine	4096	1	1
q7	32768	Joint	16384	4	0
q7	32768	Pinned	8192	4	0
q7	32768	Single long solve	16384	1	0
q7	32768	One-shot suffix	4096	8	0
q7	32768	Adapted Self-Refine	8192	1	1
q7	65536	Joint	16384	8	0
q7	65536	Pinned	8192	8	0
q7	65536	Single long solve	16384	1	0
q7	65536	One-shot suffix	16384	8	0
q7	65536	Adapted Self-Refine	4096	4	1

N MMLU-PRO EXTENSION: PROTOCOL AND COMPLETE COMPARISONS

N.1 FROZEN COHORTS, CONFIGURATIONS AND EXTRACTION

The source is `TIGER-Lab/MMLU-Pro` (Wang et al., 2024) at revision `b189ec765aa7ed75c8acfea42df31fdae71f97be`. A deterministic salted-hash split selects 50 validation and 100 test questions from each of biology, history, law and philosophy. Both subsets come from the published test pool, are disjoint from each other, and are shared across models. They are not the benchmark's official validation/test split. Exact `N/A` padding is removed as specified by the frozen preprocessor; the stored question and option projections are reused for every strategy. Validation seeds are 4101, 4102, 4103; test seeds are 5101, 5102, 5103. The two earlier MATH models retain their pinned revisions; R1-Distill-Llama-8B uses `6a6f4aa4197940add57724a7707d069478df56b1`.

Natural and suffix sampling each evaluate four allowances $c \in \{2048, 4096, 8192, 16384\}$ and six counts $N \in \{1, 2, 4, 8, 16, 32\}$. Pinned fixes $c = 8192$; depth fixes $N = 1$. Self-Refine evaluates solution allowances 4096 and 8192, trajectory counts 1, 2, 4, 8 and one or two rounds, with feedback capped at 2048. Its complete prior conversation is retained. These yield 24 natural, 24 suffix and 16 Self-Refine configurations per model. Temperature is 0.6 and top-p 0.95, with native EOS, bfloat16 and a 40960-token context. Decoding preserves the original model's chat template; Qwen3 enables thinking. The solve instruction is:

> Solve the multiple-choice question step by step. Finish with exactly one answer letter inside `\boxed{}`, for example `\boxed{A}`.

The suffix appends the same one-shot reconsideration string used elsewhere. Self-Refine feedback and revision instructions are:

> Review the previous answer and reasoning for factual or logical errors and check it against all answer options. Identify specific errors or confirm the reasoning. Do not consult any external reference.
>
> Using the original question, previous answer and feedback, produce an improved solution. Finish with exactly one answer letter inside `\boxed{}`, for example `\boxed{A}`.

The final explicit answer declaration in the returned text determines the extracted label: a boxed single letter, "Final answer:" letter, or "The answer is" letter. Malformed, ambiguous or out-of-range final declarations are invalid, rather than triggering a search for an earlier valid answer. This measures extractable-answer accuracy, including explicit answers within unfinished reasoning. Invalid samples retain their token cost but abstain; all-invalid votes are incorrect. Plurality ties use the minimum SHA-256 of the frozen protocol/question/label tuple, independent of strategy and seed. This extension's extraction and tie rules are separate from the earlier mathematics protocol.

N.2 SELECTION AND THE REGISTERED STATISTICAL FAMILY

Within each model and strategy family, selection maximizes validation accuracy under the mean input-plus-output token target, breaking ties by lower cost, fewer components, and canonical component IDs and weights. The primary selector allows a mixture of at most two configurations. Each question would choose a component independently of its answer; we evaluate the exact expectation over the generated component outcomes. This neither combines votes across components nor refits weights to test data. Unspent budget is permitted. A deterministic selector restricted to one configuration is the registered sensitivity analysis. Three targets, five families, two selectors and three models give 90 frozen selection slots.

All validation scores precede global selection and test authorization. The validation plan has 180 ten-question units and 604800 logical calls; the selected test plan has 360 units. Configuration reuse produces 115200 validation score rows and 40800 test component score rows; these are not counts of independent questions or unique generated requests. Each selected policy is charged for its expected logical input and output tokens, including repeated feedback/revision context. Neither Nc nor total experiment GPU expense is substituted for this policy cost.

Table 21: All extension strategies at the primary target, using the registered mixture selector. Qwen-7B and Llama-8B denote the DeepSeek-R1-Distill checkpoints. Costs are logical input-plus-output tokens.

Model	Strategy	Accuracy (%)	Tokens/question
Qwen3-8B	Joint	62.75	14766
Qwen3-8B	Pinned	62.45	33768
Qwen3-8B	Jointly tuned suffix	61.39	33195
Qwen3-8B	Adapted Self-Refine	62.30	33566
Qwen3-8B	Single long solve	61.83	3672
Qwen-7B	Joint	29.73	33652
Qwen-7B	Pinned	29.73	33652
Qwen-7B	Jointly tuned suffix	30.79	33364
Qwen-7B	Adapted Self-Refine	27.55	33104
Qwen-7B	Single long solve	17.67	2433
Llama-8B	Joint	41.72	32926
Llama-8B	Pinned	41.51	33015
Llama-8B	Jointly tuned suffix	36.73	31283
Llama-8B	Adapted Self-Refine	38.82	32646
Llama-8B	Single long solve	33.58	2293
Pooled	Joint	44.73	27115
Pooled	Pinned	44.56	33478
Pooled	Jointly tuned suffix	42.97	32614
Pooled	Adapted Self-Refine	42.89	33105
Pooled	Single long solve	37.69	2799

The registered primary family contains two pooled mixture comparisons at 32768: joint versus pinned and joint versus jointly tuned suffix. Accuracy is averaged over the three fixed seeds within question, then equally over categories and models. Cost ratios divide the pooled means rather than averaging individual ratios. We draw 20000 paired question-cluster bootstrap replicates within the four categories, sharing the resampled question IDs across all models and policies. There are 400 question clusters. The random generator is NumPy PCG64 with seed 20260919. For each of two contrasts, the 0.0125 accuracy-difference quantile must exceed zero and the 0.9875 cost-ratio quantile must not exceed 1.05, giving four Bonferroni-adjusted one-sided bounds. Only the suffix comparison passes; the conjunction that both comparisons pass is false. All other intervals are descriptive, conditional on the fixed selection and realized test seeds.

N.3 Absolute results and selected policies

Table 21 reports the five families for every model and the pooled mean; Table 22 reports primary mixture configurations. Qwen3 uses a longer allowance with four samples; Llama uses shorter allowances with more samples. Qwen-7B's joint and pinned mixtures are identical, giving an exact zero contrast rather than a small floating-point positive difference.

Table 22: Extension primary mixture policies. Entries are $(c, N) : w$ or $(c, N, r) : w$ for Self-Refine, with weights rounded here to four decimals. Exact rational weights are frozen in the selection record; no weights are refitted using test costs or labels.

Model	Strategy	Components and weights
Qwen3-8B	Joint	$(16384, 4) : 1.0000$
Qwen3-8B	Pinned	$(8192, 16) : 0.4466;\ (8192, 4) : 0.5534$
Qwen3-8B	Single long solve	$(16384, 1) : 1.0000$
Qwen3-8B	Jointly tuned suffix	$(16384, 2) : 0.5891;\ (4096, 32) : 0.4109$
Qwen3-8B	Adapted Self-Refine	$(4096, 1, 1) : 0.4367;\ (8192, 4, 1) : 0.5633$
Qwen-7B	Joint	$(8192, 16) : 0.8281;\ (8192, 4) : 0.1719$
Qwen-7B	Pinned	$(8192, 16) : 0.8281;\ (8192, 4) : 0.1719$
Qwen-7B	Single long solve	$(8192, 1) : 1.0000$
Qwen-7B	Jointly tuned suffix	$(16384, 16) : 0.5117;\ (4096, 16) : 0.4883$
Qwen-7B	Adapted Self-Refine	$(4096, 2, 1) : 0.1530;\ (4096, 4, 1) : 0.8470$
Llama-8B	Joint	$(2048, 16) : 0.8195;\ (4096, 32) : 0.1805$
Llama-8B	Pinned	$(8192, 16) : 0.8943;\ (8192, 8) : 0.1057$
Llama-8B	Single long solve	$(16384, 1) : 1.0000$
Llama-8B	Jointly tuned suffix	$(8192, 16) : 0.6990;\ (8192, 32) : 0.3010$
Llama-8B	Adapted Self-Refine	$(4096, 4, 1) : 0.9145;\ (8192, 8, 1) : 0.0855$

Table 23: Complete extension contrasts at target 16384 with the mixture selector. Direction is joint minus comparator accuracy and joint/comparator cost. All intervals are descriptive 95% question-cluster intervals; only the two pooled mixture contrasts at 32768 have registered primary decisions.

Model	Comparator	Δ accuracy [95% CI], pp	Cost ratio [95% CI]
Qwen3-8B	Pinned	$+0.69$ $[-0.77, +2.18]$	0.875 $[0.864, 0.888]$
Qwen3-8B	Jointly tuned suffix	$+1.79$ $[-0.14, +3.71]$	0.889 $[0.871, 0.906]$
Qwen3-8B	Adapted Self-Refine	$+0.80$ $[-0.91, +2.50]$	0.881 $[0.863, 0.901]$
Qwen3-8B	Single long solve	$+0.92$ $[-0.75, +2.58]$	4.021 $[3.959, 4.084]$
Qwen-7B	Pinned	$+0.00$ $[+0.00, +0.00]$	1.000 $[1.000, 1.000]$
Qwen-7B	Jointly tuned suffix	-1.36 $[-3.55, +0.79]$	1.016 $[0.994, 1.039]$
Qwen-7B	Adapted Self-Refine	$+2.54$ $[+0.46, +4.61]$	1.016 $[0.986, 1.046]$
Qwen-7B	Single long solve	$+10.04$ $[+7.87, +12.26]$	6.934 $[6.725, 7.151]$
Llama-8B	Pinned	$+0.49$ $[-1.55, +2.55]$	1.002 $[0.963, 1.042]$
Llama-8B	Jointly tuned suffix	$+4.40$ $[+1.84, +6.93]$	1.045 $[0.985, 1.106]$
Llama-8B	Adapted Self-Refine	$+4.63$ $[+2.15, +7.15]$	1.004 $[0.982, 1.027]$
Llama-8B	Single long solve	$+6.25$ $[+3.52, +9.02]$	7.173 $[6.698, 7.686]$
Pooled	Pinned	$+0.39$ $[-0.48, +1.26]$	0.959 $[0.947, 0.970]$
Pooled	Jointly tuned suffix	$+1.61$ $[+0.29, +2.91]$	0.982 $[0.966, 0.998]$
Pooled	Adapted Self-Refine	$+2.66$ $[+1.34, +3.96]$	0.967 $[0.956, 0.978]$
Pooled	Single long solve	$+5.73$ $[+4.46, +7.01]$	5.725 $[5.598, 5.857]$

Table 24: Complete extension contrasts at target 32768 with the mixture selector. Direction is joint minus comparator accuracy and joint/comparator cost. All intervals are descriptive 95% question-cluster intervals; only the two pooled mixture contrasts at 32768 have registered primary decisions.

Model	Comparator	Δ accuracy [95% CI], pp	Cost ratio [95% CI]
Qwen3-8B	Pinned	$+0.30$ $[-1.09, +1.72]$	0.437 $[0.432, 0.444]$
Qwen3-8B	Jointly tuned suffix	$+1.36$ $[-0.44, +3.14]$	0.445 $[0.435, 0.455]$
Qwen3-8B	Adapted Self-Refine	$+0.45$ $[-0.82, +1.71]$	0.440 $[0.433, 0.447]$
Qwen3-8B	Single long solve	$+0.92$ $[-0.75, +2.58]$	4.021 $[3.959, 4.084]$
Qwen-7B	Pinned	$+0.00$ $[+0.00, +0.00]$	1.000 $[1.000, 1.000]$
Qwen-7B	Jointly tuned suffix	-1.06 $[-2.91, +0.79]$	1.009 $[0.989, 1.029]$
Qwen-7B	Adapted Self-Refine	$+2.18$ $[+0.16, +4.20]$	1.017 $[0.988, 1.045]$
Qwen-7B	Single long solve	$+12.07$ $[+9.45, +14.74]$	13.833 $[13.385, 14.290]$
Llama-8B	Pinned	$+0.21$ $[-1.58, +2.03]$	0.997 $[0.965, 1.031]$
Llama-8B	Jointly tuned suffix	$+4.98$ $[+2.42, +7.54]$	1.053 $[1.000, 1.107]$
Llama-8B	Adapted Self-Refine	$+2.90$ $[+0.89, +4.89]$	1.009 $[0.991, 1.026]$
Llama-8B	Single long solve	$+8.14$ $[+5.32, +10.97]$	14.360 $[13.471, 15.317]$
Pooled	Pinned	$+0.17$ $[-0.60, +0.94]$	0.810 $[0.801, 0.819]$
Pooled	Jointly tuned suffix	$+1.76$ $[+0.53, +3.01]$	0.831 $[0.818, 0.845]$
Pooled	Adapted Self-Refine	$+1.84$ $[+0.78, +2.90]$	0.819 $[0.809, 0.829]$
Pooled	Single long solve	$+7.04$ $[+5.66, +8.44]$	9.686 $[9.465, 9.917]$

N.4 ALL MODELS, TARGETS AND SELECTORS

The six following tables retain all 96 registered model/pooled contrasts. The pooled mixture selector favors joint over the suffix in both accuracy and cost at all three targets. Single-configuration selection at 32768 gives $+1.61$ points at 8.09% higher cost; versus pinned it gives $+0.50$ points at 21.96% higher cost. Thus the main simultaneous accuracy/cost improvement does not extend unchanged to deterministic selection. At the primary mixture target, pooled cost savings are mainly contributed by Qwen3, whereas the largest accuracy gain is on Llama; the Qwen-7B suffix is more accurate in point estimate. A pooled result is not per-model dominance.

Table 25: Complete extension contrasts at target 65536 with the mixture selector. Direction is joint minus comparator accuracy and joint/comparator cost. All intervals are descriptive 95% question-cluster intervals; only the two pooled mixture contrasts at 32768 have registered primary decisions.

Model	Comparator	Δ accuracy [95% CI], pp	Cost ratio [95% CI]
Qwen3-8B	Pinned	-0.25 $[-1.83, +1.33]$	0.256 $[0.252, 0.259]$
Qwen3-8B	Jointly tuned suffix	$+0.52$ $[-1.55, +2.54]$	0.223 $[0.218, 0.228]$
Qwen3-8B	Adapted Self-Refine	$+0.08$ $[-1.25, +1.42]$	0.291 $[0.286, 0.295]$
Qwen3-8B	Single long solve	$+0.92$ $[-0.75, +2.58]$	4.021 $[3.959, 4.084]$
Qwen-7B	Pinned	$+0.00$ $[+0.00, +0.00]$	1.000 $[1.000, 1.000]$
Qwen-7B	Jointly tuned suffix	-0.50 $[-2.67, +1.67]$	1.019 $[0.990, 1.050]$
Qwen-7B	Adapted Self-Refine	$+0.66$ $[-1.24, +2.57]$	0.582 $[0.566, 0.598]$
Qwen-7B	Single long solve	$+12.67$ $[+9.83, +15.58]$	15.878 $[15.359, 16.408]$
Llama-8B	Pinned	$+0.58$ $[-1.25, +2.42]$	1.806 $[1.765, 1.847]$
Llama-8B	Jointly tuned suffix	$+5.25$ $[+2.42, +8.17]$	1.306 $[1.252, 1.362]$
Llama-8B	Adapted Self-Refine	$+1.00$ $[-1.08, +3.08]$	0.976 $[0.957, 0.994]$
Llama-8B	Single long solve	$+8.75$ $[+5.92, +11.58]$	27.460 $[25.944, 29.086]$
Pooled	Pinned	$+0.11$ $[-0.72, +0.92]$	0.886 $[0.876, 0.897]$
Pooled	Jointly tuned suffix	$+1.76$ $[+0.34, +3.15]$	0.763 $[0.750, 0.776]$
Pooled	Adapted Self-Refine	$+0.58$ $[-0.48, +1.61]$	0.640 $[0.633, 0.647]$
Pooled	Single long solve	$+7.44$ $[+6.00, +8.92]$	13.856 $[13.533, 14.191]$

Table 26: Complete extension contrasts at target 16384 with the single-configuration selector. Direction is joint minus comparator accuracy and joint/comparator cost. All intervals are descriptive 95% question-cluster intervals; only the two pooled mixture contrasts at 32768 have registered primary decisions.

Model	Comparator	Δ accuracy [95% CI], pp	Cost ratio [95% CI]
Qwen3-8B	Pinned	$+0.75$ $[-0.75, +2.25]$	1.026 $[1.012, 1.042]$
Qwen3-8B	Jointly tuned suffix	$+2.08$ $[-0.00, +4.17]$	2.934 $[2.865, 3.003]$
Qwen3-8B	Adapted Self-Refine	$+0.92$ $[-1.00, +2.83]$	1.306 $[1.271, 1.343]$
Qwen3-8B	Single long solve	$+0.92$ $[-0.75, +2.58]$	4.021 $[3.959, 4.084]$
Qwen-7B	Pinned	$+0.00$ $[+0.00, +0.00]$	1.000 $[1.000, 1.000]$
Qwen-7B	Jointly tuned suffix	-1.92 $[-4.42, +0.50]$	0.673 $[0.656, 0.691]$
Qwen-7B	Adapted Self-Refine	$+6.58$ $[+4.25, +9.00]$	1.073 $[1.036, 1.111]$
Qwen-7B	Single long solve	$+9.17$ $[+7.00, +11.33]$	3.977 $[3.864, 4.094]$
Llama-8B	Pinned	$+1.17$ $[-1.42, +3.75]$	1.515 $[1.453, 1.579]$
Llama-8B	Jointly tuned suffix	$+4.33$ $[+1.67, +7.00]$	1.093 $[1.028, 1.160]$
Llama-8B	Adapted Self-Refine	$+4.42$ $[+1.75, +7.17]$	0.881 $[0.861, 0.902]$
Llama-8B	Single long solve	$+5.67$ $[+2.92, +8.50]$	5.739 $[5.358, 6.151]$
Pooled	Pinned	$+0.64$ $[-0.39, +1.64]$	1.148 $[1.134, 1.162]$
Pooled	Jointly tuned suffix	$+1.50$ $[+0.08, +2.92]$	1.196 $[1.167, 1.225]$
Pooled	Adapted Self-Refine	$+3.97$ $[+2.56, +5.39]$	1.067 $[1.052, 1.081]$
Pooled	Single long solve	$+5.25$ $[+3.97, +6.50]$	4.477 $[4.376, 4.581]$

Table 27: Complete extension contrasts at target 32768 with the single-configuration selector. Direction is joint minus comparator accuracy and joint/comparator cost. All intervals are descriptive 95% question-cluster intervals; only the two pooled mixture contrasts at 32768 have registered primary decisions.

Model	Comparator	Δ accuracy [95% CI], pp	Cost ratio [95% CI]
Qwen3-8B	Pinned	$+0.75$ $[-0.75, +2.25]$	1.026 $[1.012, 1.042]$
Qwen3-8B	Jointly tuned suffix	$+2.08$ $[-0.00, +4.17]$	2.934 $[2.865, 3.003]$
Qwen3-8B	Adapted Self-Refine	$+0.92$ $[-1.00, +2.83]$	1.306 $[1.271, 1.343]$
Qwen3-8B	Single long solve	$+0.92$ $[-0.75, +2.58]$	4.021 $[3.959, 4.084]$
Qwen-7B	Pinned	-1.33 $[-3.42, +0.75]$	1.098 $[1.069, 1.127]$
Qwen-7B	Jointly tuned suffix	-2.25 $[-4.75, +0.25]$	0.742 $[0.714, 0.770]$
Qwen-7B	Adapted Self-Refine	$+2.50$ $[-0.17, +5.17]$	1.186 $[1.133, 1.240]$
Qwen-7B	Single long solve	$+10.83$ $[+8.25, +13.50]$	8.725 $[8.367, 9.091]$
Llama-8B	Pinned	$+2.08$ $[+0.00, +4.17]$	1.516 $[1.458, 1.576]$
Llama-8B	Jointly tuned suffix	$+5.00$ $[+2.25, +7.67]$	1.096 $[1.036, 1.159]$
Llama-8B	Adapted Self-Refine	$+3.00$ $[+0.75, +5.25]$	0.887 $[0.868, 0.905]$
Llama-8B	Single long solve	$+8.00$ $[+5.08, +10.92]$	11.474 $[10.718, 12.292]$
Pooled	Pinned	$+0.50$ $[-0.61, +1.58]$	1.220 $[1.202, 1.238]$
Pooled	Jointly tuned suffix	$+1.61$ $[+0.17, +3.06]$	1.081 $[1.057, 1.106]$
Pooled	Adapted Self-Refine	$+2.14$ $[+0.72, +3.53]$	1.058 $[1.042, 1.075]$
Pooled	Single long solve	$+6.58$ $[+5.19, +7.97]$	7.419 $[7.238, 7.611]$

Table 28: Complete extension contrasts at target 65536 with the single-configuration selector. Direction is joint minus comparator accuracy and joint/comparator cost. All intervals are descriptive 95% question-cluster intervals; only the two pooled mixture contrasts at 32768 have registered primary decisions.

Model	Comparator	Δ accuracy [95% CI], pp	Cost ratio [95% CI]
Qwen3-8B	Pinned	-0.25 $[-1.83, +1.33]$	0.256 $[0.252, 0.259]$
Qwen3-8B	Jointly tuned suffix	$+0.83$ $[-1.08, +2.83]$	0.370 $[0.364, 0.377]$
Qwen3-8B	Adapted Self-Refine	$+0.08$ $[-1.25, +1.42]$	0.291 $[0.286, 0.295]$
Qwen3-8B	Single long solve	$+0.92$ $[-0.75, +2.58]$	4.021 $[3.959, 4.084]$
Qwen-7B	Pinned	$+0.00$ $[+0.00, +0.00]$	1.000 $[1.000, 1.000]$
Qwen-7B	Jointly tuned suffix	-0.50 $[-2.67, +1.67]$	1.019 $[0.990, 1.050]$
Qwen-7B	Adapted Self-Refine	$+2.50$ $[+0.25, +4.75]$	1.077 $[1.047, 1.108]$
Qwen-7B	Single long solve	$+12.67$ $[+9.83, +15.58]$	15.878 $[15.359, 16.408]$
Llama-8B	Pinned	$+0.58$ $[-1.25, +2.42]$	1.806 $[1.765, 1.847]$
Llama-8B	Jointly tuned suffix	$+5.25$ $[+2.42, +8.17]$	1.306 $[1.252, 1.362]$
Llama-8B	Adapted Self-Refine	$+1.00$ $[-1.08, +3.08]$	0.976 $[0.957, 0.994]$
Llama-8B	Single long solve	$+8.75$ $[+5.92, +11.58]$	27.460 $[25.944, 29.086]$
Pooled	Pinned	$+0.11$ $[-0.72, +0.92]$	0.886 $[0.876, 0.897]$
Pooled	Jointly tuned suffix	$+1.86$ $[+0.44, +3.25]$	0.924 $[0.903, 0.944]$
Pooled	Adapted Self-Refine	$+1.19$ $[+0.06, +2.33]$	0.770 $[0.760, 0.779]$
Pooled	Single long solve	$+7.44$ $[+6.00, +8.92]$	13.856 $[13.533, 14.191]$

Table 29: Per-test-seed pooled extension contrasts at target 32768 under the mixture selector. These are descriptive point estimates, not new hypothesis tests.

Seed	Comparator	Δ accuracy (pp)	Cost ratio
5101	Pinned	-0.37	0.811
5101	Jointly tuned suffix	$+1.64$	0.828
5101	Adapted Self-Refine	$+1.49$	0.817
5101	Single long solve	$+6.89$	9.644
5102	Pinned	$+0.89$	0.809
5102	Jointly tuned suffix	$+2.57$	0.832
5102	Adapted Self-Refine	$+2.83$	0.815
5102	Single long solve	$+6.59$	9.660
5103	Pinned	-0.01	0.810
5103	Jointly tuned suffix	$+1.08$	0.835
5103	Adapted Self-Refine	$+1.21$	0.825
5103	Single long solve	$+7.65$	9.755

N.5 Seed sensitivity and numerical reproduction

Each individual test seed favors the joint mixture over the suffix in pooled accuracy and token cost (Table 29). Omitting one seed at a time gives accuracy differences from $+1.36$ to $+2.10$ points and cost ratios from 0.830 to 0.833. The pinned comparison's leave-one-seed-out accuracy differences range from -0.19 to $+0.44$ points; its accuracy superiority is not established. These checks are descriptive, not additional independent replications or a seed-resampling confidence interval.

The score-level reproduction reconstructs validation summaries and all 90 selection slots, then recomputes the primary bootstrap bounds and decisions. The returned configuration scores reproduce those quantities exactly; platform-level floating-point differences in secondary displays are below 10^{-10}. This check is distinct from regrading raw model outputs or auditing the experiment's unique-generation and allocated GPU costs. The accompanying numerical package states that coverage explicitly.

O SEPARATELY REGISTERED 32B CONFIRMATION

O.1 DESIGN AND PRIMARY RESULTS

DeepSeek-R1-Distill-Qwen-32B uses the original bfloat16 checkpoint at revision `711ad2ea6aa40cfca18895e8aca02ab92df1a746`. The study reuses the exact 200 validation and 400 test questions from the four MMLU-Pro categories in Appendix N, with the same prompts, extraction, sampling settings and 64-configuration search. Validation seeds are 4101, 4102, 4103 and test seeds are 5101, 5102, 5103. Its own protocol identifier determines generation seeds and plurality tie-breaking. Validation registration precedes generation; independently selected policies and a separate test authorization precede test generation. This is a larger-model check on shared questions, not a new question cohort; it is neither pooled with the three-model study nor added to its primary testing denominator. The historical 32B checkpoint was not revision-pinned.

The five families, three targets and two selectors give 30 selection slots. All selected policies satisfy their validation mean-token ceilings and retain their complete selected-component scores. The study has 60 ten-question validation units (201600 logical calls), 120 test units, 38400 validation configuration score rows and 12000 selected test configuration score rows. Inference uses same-node tensor parallelism: two identical 48GB GPUs (L20 or RTX5880), or four identical 24GB GPUs (A30 or RTX3090), with disjoint GPU groups and a 40960-token context. Within each question/seed shard, all strategies share the assigned hardware.

Table 30: 32B accuracy and measured mean input-plus-output token cost at the primary 32768 selection target. All five strategy families use validation-frozen policies.

Strategy	Accuracy (%)	Tokens/question
Joint natural	64.33	28563.35
Pinned natural	64.33	28563.35
Tuned suffix	61.83	8488.74
Adapted Self-Refine	63.36	33060.27
Single long solve	61.58	1810.70

The primary target is 32768 and the primary selector allows at most two configurations with frozen, outcome-independent weights. Question/seed scores are component expectations, not pooled votes. Unspent budget is permitted. Means average seeds within question and weight the four categories equally. There are 400 question clusters. The registered 20000-draw category-stratified paired bootstrap uses PCG64 seed 20260919. For each of two primary comparisons, the accuracy-difference 0.0125 quantile must exceed zero and the cost-ratio 0.9875 quantile must be at most 1.05. These four one-sided bounds target a nominal familywise level of 0.05, conditional on fixed validation selection and realized test seeds.

Table 31: Both registered 32B primary decisions. Accuracy differences are joint minus comparator; ratios are joint/comparator cost. Bounds use the registered adjusted quantiles.

Comparator	Accuracy lower (pp)	Cost-ratio upper	Joint criterion
Pinned	+0.000	1.0000	No
Suffix	+0.167	3.4639	No

Joint and pinned both select the same natural $(c, N) = (8192, 16)$ policy, referencing the same generated records, so their exact tie is a policy identity rather than an independent equivalence test. Joint's advantage over the tuned suffix is $+2.50$ points at $3.365\times$ the actual cost. Neither of the two primary comparisons meets the joint criterion. Against adapted Self-Refine, the secondary point estimate is $+0.97$ points at 13.60% lower cost, with an accuracy-difference interval spanning zero.

O.2 Complete comparisons across targets and selectors

Table 32 retains all 24 distinct comparisons: four comparators, three targets and two selectors. The statistics file's `POOLED` entries duplicate this single-model analysis and are not additional tests. All secondary intervals are descriptive. At 16384 the joint-versus-pinned accuracy point estimate is negative; at the two larger targets the selected policies coincide. A larger validation search space does not guarantee a held-out gain. For the two primary comparisons, single-configuration selection yields the same policies as the primary selector, so their conclusions do not depend on mixture evaluation.

Table 32: All 24 distinct 32B comparisons. Entries are joint minus comparator accuracy and joint/comparator cost; brackets give descriptive 95% question-cluster intervals. Only the two mixture comparisons against pinned and suffix at 32768 are primary.

Comparator	Δ accuracy [95% CI], pp	Cost ratio [95% CI]
Mixture, target 16384		
Pinned	-1.07 $[-2.17, +0.02]$	1.001 $[0.993, 1.009]$
Suffix	$+1.29$ $[-0.63, +3.25]$	1.977 $[1.926, 2.030]$
Self-Refine	$+1.52$ $[+0.01, +3.05]$	1.022 $[0.996, 1.047]$
Single solve	$+1.54$ $[-0.60, +3.68]$	9.271 $[9.047, 9.502]$
Mixture, target 32768		
Pinned	$+0.00$ $[+0.00, +0.00]$	1.000 $[1.000, 1.000]$
Suffix	$+2.50$ $[+0.42, +4.67]$	3.365 $[3.282, 3.451]$
Self-Refine	$+0.97$ $[-0.60, +2.59]$	0.864 $[0.842, 0.886]$
Single solve	$+2.75$ $[+0.42, +5.08]$	15.775 $[15.397, 16.169]$
Mixture, target 65536		
Pinned	$+0.00$ $[+0.00, +0.00]$	1.000 $[1.000, 1.000]$
Suffix	$+2.50$ $[+0.42, +4.67]$	3.365 $[3.282, 3.451]$
Self-Refine	$+0.08$ $[-1.50, +1.58]$	0.537 $[0.524, 0.551]$
Single solve	$+2.75$ $[+0.42, +5.08]$	15.775 $[15.397, 16.169]$
Single configuration, target 16384		
Pinned	-1.42 $[-2.92, +0.00]$	1.053 $[1.029, 1.079]$
Suffix	$+0.92$ $[-1.17, +3.00]$	1.775 $[1.712, 1.841]$
Self-Refine	$+3.67$ $[+1.33, +6.00]$	1.086 $[1.051, 1.122]$
Single solve	$+1.17$ $[-1.17, +3.50]$	8.322 $[8.072, 8.574]$
Single configuration, target 32768		
Pinned	$+0.00$ $[+0.00, +0.00]$	1.000 $[1.000, 1.000]$
Suffix	$+2.50$ $[+0.42, +4.67]$	3.365 $[3.282, 3.451]$
Self-Refine	$+1.25$ $[-0.42, +3.00]$	1.070 $[1.043, 1.098]$
Single solve	$+2.75$ $[+0.42, +5.08]$	15.775 $[15.397, 16.169]$
Single configuration, target 65536		
Pinned	$+0.00$ $[+0.00, +0.00]$	1.000 $[1.000, 1.000]$
Suffix	$+2.50$ $[+0.42, +4.67]$	3.365 $[3.282, 3.451]$
Self-Refine	$+0.08$ $[-1.50, +1.58]$	0.537 $[0.524, 0.551]$
Single solve	$+2.75$ $[+0.42, +5.08]$	15.775 $[15.397, 16.169]$

O.3 SELECTED SETTINGS, SEED CHECKS AND REPRODUCTION

Table 33: Primary-target 32B settings. Coordinates are (c, N) for sampling and (c, N, r) for Self-Refine; values after colons are mixture weights. The score package retains exact rational weights.

Strategy	Selected components and weights
Joint natural	$(8192, 16) : 1.000000$
Pinned natural	$(8192, 16) : 1.000000$
Tuned suffix	$(4096, 8) : 1.000000$
	$(4096, 4, 1) : 0.759592$
Adapted Self-Refine	$(4096, 8, 1) : 0.240408$
Single long solve	$(8192, 1) : 1.000000$

At the primary target, the selected natural policy's validation accuracy and cost are 65.67% and 27875.38 tokens, compared with 63.00% and 8476.88 tokens for the suffix policy. The selector maximizes validation accuracy under a cost ceiling, resolving ties by lower cost; it does not force a family to spend the target. Consequently the shared target is not an equal-cost comparison. Table 34 reports the primary suffix comparison's seed and category sensitivity; every omitted-seed estimate remains more accurate and more costly. Pinned contrasts remain exact ties throughout because the policies are identical.

Table 34: Primary-target joint versus suffix sensitivity. These are descriptive point estimates, not independent hypothesis tests. Each category has 100 question clusters.

Subset	Δ accuracy (pp)	Cost ratio
Seed 5101	+3.50	3.279
Seed 5102	+2.75	3.381
Seed 5103	+1.25	3.439
Omit seed 5101	+2.000	3.409
Omit seed 5102	+2.375	3.357
Omit seed 5103	+3.125	3.329
Biology	+2.000	3.194
History	-0.333	3.227
Law	+3.667	3.169
Philosophy	+4.667	4.052

The score-level CPU package rebuilds the validation summary from all 38400 configuration scores, reproduces all 30 frozen selections and recomputes the 20000-draw primary bounds. Both primary decisions reproduce exactly; secondary floating-point displays agree within 10^{-10}. The package retains all targets and selectors, including the different Self-Refine cost trade-offs. Its provenance distinguishes the new model generations from the shared cohort. These checks reproduce scores to selection and statistics, not raw-output extraction, per-generation configuration receipts or allocated GPU-time accounting.

P SUPPORTING PROMPT DIAGNOSTICS

P.1 DIRECTION AND ALTERNATIVE WORDINGS

Direction is a separate diagnostic. Let R be the count of natural-wrong/suffix-correct pairs divided by the reverse count. One open-ended cell has R within the observed multiple-choice range of 0.89–1.06, and two have still lower ratios. Directional effects therefore vary within each format. The main suffix and two alternatives on one AIME cell give $D_1 = 18.5\%, 18.8\%, 16.2\%$ but $R = 0.89, 0.88, 1.29$ (Appendix K). A four-condition registration on two additional task families met its original literal criteria. A contemporaneous note discusses near-zero effects, but its blindness is unresolved (Appendix B).

P.2 RE-RENDERING THE SAME QUESTIONS

We re-render MATH-500 and AIME with four labelled answer options for two models, retaining the questions and correct answers while adding distractors. Discordance rises in all four comparisons, by factors of 1.5–6.0 (Appendix G). An audit against the original answer keys finds a unique correct option in all 160 items. Six MATH-500 items repeat a wrong distractor; excluding those items from both renderings leaves 94 paired questions and discordance increases of $5.87\times$ and $4.50\times$. Thus the increase persists after removing duplicated distractors. The intervention combines presentation, answer cues and label extraction. Baseline accuracy rises substantially for Qwen3-8B: 76% to 94% on MATH-500 and 40% to 63.3% on AIME, illustrating the accompanying change in effective difficulty.

The registered thresholds were more demanding than a positive direction: one of six predictions held, each requiring a specified effect in both models. These counts retain the full 100- and 60-question cohorts; the 94-question analysis is a post-hoc sensitivity check. A natural–natural control within each rendering would further separate sampling variability from excess prompt discordance. Appendix F reports the accompanying aggregation comparisons.

Length has a separate unit. The suffix produces fewer generated *characters* in fifteen of sixteen cells, with the largest reported reduction about 46%. Appendix C retains the family summaries and weighting. This is a text-length measurement, distinct from realized token use.

Reconciling the comparisons. On Qwen3-8B/AIME, changing the natural baseline from its pinned-row maximum to its grid maximum raises it from 41.7% to 70.0%, a 28.3-point difference; the gap to the suffix peak is then 1.7 points. The identity $30.0 = 28.3 + 1.7$ compares optima at different allocation coordinates. The fixed-(N, c) ablation above separately measures the prompt's effect.

Table 35: All recorded rounds. The *sensitivity* column excludes the same two round-3 cases as the earlier post-hoc reading; it is not a replacement registration. Zero credit in the forcing rows includes unmet gates and void downstream criteria, not three valid negative efficacy tests per row. The format-item audit leaves rounds 5–6 unchanged. Original criteria for rounds 7–8 are reproduced in Appendix R.

Round	Registered target	Literal	Sensitivity
1	Binary crossover	4/6	4/6
2	Three margin bands	2/3	2/3
3	Convergence, activity, fragmentation gates	3/6	2/4
4	Answer-space hypothesis	4/4	4/4
5	MATH-500 format intervention	1/3	1/3
6	AIME format intervention	0/3	0/3
7	Forcing at 8192: gate failed	0/3	0/3
8	Forcing at 16384: gate failed	0/3	0/3
	Rounds 1–6	14/25	13/23
	Historical all-condition count	14/31	13/29

Q HISTORICAL PROSPECTIVE DIAGNOSTICS

A useful allocation rule would predict where additional allowance pays from observations cheaper than a full grid. We test registered rules and feature-based regressions as candidates for this role.

Registered attempts, separated by purpose. Table 35 retains all eight recorded rounds. Only the first three predict the margin between axes, with 9 of 15 literal criteria met. Rounds 4–6 address prompt and format hypotheses; rounds 7–8 concern forcing engagement and contain gated efficacy criteria that never became valid tests. Their combined count is an audit trail, not a classifier accuracy.

The literal and sensitivity counts preserve two protocol distinctions: a round-1 entry followed its run's start, and the round-3 sensitivity reading removes two empirically near-saturated cells, one hit and one miss, after an omitted scope condition was identified. The latter is a post-hoc reading. Appendix B records these cases alongside the original criteria in Appendix R.

Predictive models and practical consequence. Twelve low-allowance feature sets, including fragmentation, vote convergence, single-sample accuracy and sequential length growth, were evaluated by leave-one-cell-out regression. None improved on the mean-prediction baseline in the recorded comparison. Features that use suffix probes include their additional generation cost.

The distributional account suggests a reason extrapolation can be difficult: additional allowance can change answer probabilities beyond those observed in a low-allowance bank. Direct grid measurements supply the calibration evidence here; developing a cheaper policy is a prospective learning problem.

R SELECTED REGISTRATION EXCERPTS

These source excerpts contain three entries from round 3 and the criterion blocks for rounds 4–8, with line wrapping introduced for typesetting. Rounds 1–2 are not reproduced here. The returned registry supplied the forcing blocks for rounds 7–8; its source hash is recorded in the reproduction package. Queue snapshots alone do not establish absence of prior data, and the round-4 amendment below lists four result files whose content is not documented in this excerpt. Original wording, including causal claims, older metric names and the wave-9 explanation of engagement, is retained as historical record. The current analysis and measured engagement rates are given in Section 5 and Appendices Q and J.

ROUND 3: THREE GATE-PROCEDURE ENTRIES

```
# Wave-3 machine-generated registration log

Gate parameters frozen 2026-08-09 19:30 (+08) in
prep/blind_predictions_wave3.md.
Entries below are appended by auto_register.py the moment a cell's
parallel axis and
2-point probe are complete, BEFORE its full sequential ladder is
submitted.

### X3: deepseek-ai__DeepSeek-R1-Distill-Llama-8B__T4_bbh_ruin
- registered (machine): **2026-08-10T05:31:25+08:00** | full
sequential ladder not yet
    submitted at this instant
- parallel: nq=250, vote@16=70.0%, vote@64=72.4%
- gate1 vote-convergence: g1=+2.4pp (FAIL -> not-yet-converged)
- gate2 sequential activity: len2048=1886 len8192=2054 ratio=1.09x
(FAIL -> inert)
- gate3 fragmentation: frag@8=2.46
- **PREDICTION: margin <= 2pp**  (decided by gate1
NOT-YET-CONVERGED)

### X6: deepseek-ai__DeepSeek-R1-Distill-Llama-8B__T5_v2_raw
- registered (machine): **2026-08-10T06:01:26+08:00** | full
sequential ladder not yet
    submitted at this instant
- parallel: nq=64, vote@16=28.1%, vote@64=34.4%
- gate1 vote-convergence: g1=+6.2pp (FAIL -> not-yet-converged)
- gate2 sequential activity: len2048=4444 len8192=6803 ratio=1.53x
(FAIL -> inert)
- gate3 fragmentation: frag@8=5.36
- **PREDICTION: margin <= 2pp**  (decided by gate1
NOT-YET-CONVERGED)

### X1: Qwen__Qwen3-8B__T5_adversarial_raw
- registered (machine): **2026-08-11T10:43:06+08:00** | full
sequential ladder not yet
    submitted at this instant
- parallel: nq=160, vote@16=95.0%, vote@64=95.0%
- gate1 vote-convergence: g1=+0.0pp (pass)
- gate2 sequential activity: len2048=4573 len8192=7870 ratio=1.72x
(FAIL -> inert)
- gate3 fragmentation: frag@8=1.14
- **PREDICTION: margin <= 2pp**  (decided by gate2 INERT-SEQ)
```

ROUNDS 4–6 (ANSWER FORMAT, FORMAT FLIP)

```
=== WAVE-4 FORMAT HYPOTHESIS, registered 2026-08-10T12:13:49Z
(machine UTC) ===
Hypothesis: the sign of the reconsideration-suffix effect is set
by the answer-space
structure (closed multiple-choice vs open-ended), not by task
difficulty or domain.
Frozen criteria (matched cap 8192, matched N=8, same questions,
majority vote):
  Y1  BBH ruin_names (MC), pooled over Qwen-7B + Qwen3-8B : delta
acc <= 0
  Y2  MATH-500 (open-ended), pooled over Qwen-7B + Qwen3-8B :
delta acc > 0
  Y3  flip ratio (wrong->right / right->wrong) BBH < MATH-500, and
BBH in [0.80, 1.15]
  Y4  mean generated length shorter with suffix on BOTH families
At registration time NO job for these four cells has been
submitted and no data exists.
squeue at registration:
    1725826           w3p_x1_q3_t5  R   19:49:23
    1727793           cs_m1_c16384  R   11:38:48
    1729713        cs_m6_c16384_c2 PD       0:00
    1727798            cs_m6_c4096  R    8:44:16
    1727796           cs_m6_c16384  R   11:38:47

--- amendment 2026-08-10T12:18:11Z (machine UTC), still before any
wave-4 result file exists
     ---
Pre-disclosed weakness: Y1 is judged on the point estimate, so a
truly-null effect makes
it a coin flip (same defect class as the round-1 P4 protocol
hole). The criterion is NOT
changed. Interpretation rule, frozen now:
  delta significantly negative (CI upper < 0)            ->
supports H
  not significant, churn high and ratio in [0.80,1.15]   ->
supports the mechanism only
  delta significantly positive (CI lower > 0)            ->
falsifies H
  not significant and churn low (< 10%)                  -> NO
INFORMATION, cell excluded
                                                            from
the evidential reading
                                                            (still
reported, still counted)
Existing w4 output files at amendment time:
4 jsonl files

=== WAVE-6 FORMAT-FLIP (causal test), registered
2026-08-12T01:52:53Z (machine UTC) ===
Design: the SAME 100 MATH-500 questions, same correct answers,
re-rendered as
four-way multiple choice (distractors drawn from other questions
answers, type
matched where the type group allows). Only the response format
changes; content
is held fixed. Models Qwen-7B and Qwen3-8B, matched allowance
8192, N=8.
Observed on the OPEN-ENDED rendering of these same questions
(already measured):
  rewrite volume  Qwen-7B 7.9%   Qwen3-8B 5.0%
```

```
Observed on the two genuine multiple-choice families:  31.1% -
38.0%
FROZEN PREDICTIONS:
  Z1  rewrite volume on the MC rendering > 25% for BOTH models
  Z2  rewrite volume on the MC rendering is at least 3x its
open-ended value,
      for BOTH models
  Z3  the flip ratio (wrong->right / right->wrong) falls on the MC
rendering
      relative to the open-ended rendering, for BOTH models
Interpretation rule, frozen now: Z1/Z2 test whether ANSWER FORMAT
ALONE drives
rewrite volume, since content is identical. If Z1/Z2 fail, the
cell-level volume
separation reported in the paper is confounded with task content
and the claim
must be weakened to a correlation. Z3 is secondary; the ratio is
already known
to be noisy at cell level and a Z3 miss does not by itself refute
the volume claim.
At registration no job for these four cells has been submitted and
no data exists.
squeue at registration:
    1730059     w5_06_q3_t7_c16384  R 1-10:56:56
    1729713        cs_m6_c16384_c2  R 1-12:26:39

=== WAVE-7 FORMAT-FLIP REPLICATION on AIME, registered
2026-08-12T02:58:59Z (machine UTC) ==
    =
Same construction as wave-6 but on a second content set: the 60
AIME questions,
same correct answers, re-rendered as four-way multiple choice.
AIME answers are
all integers in 0-999, so distractors are drawn from the other
questions
answers and are type-matched by construction (no fallback needed,
unlike the
MATH-500 rendering where 5 of 100 questions used cross-type
distractors).
Models Qwen-7B and Qwen3-8B, matched allowance 8192, N=8.
Observed on the OPEN-ENDED rendering of these same questions:
  rewrite volume  Qwen-7B 18.5%   Qwen3-8B 13.5%
  flip ratio      Qwen-7B 0.89    Qwen3-8B 2.10
FROZEN PREDICTIONS:
  W1  rewrite volume on the MC rendering > 25% for BOTH models
  W2  rewrite volume on the MC rendering is at least 2x its
open-ended value,
      for BOTH models
  W3  the flip ratio on the MC rendering stays above 1.2 for BOTH
models
      (the two-factor reading: format drives VOLUME, while whether
rewriting
      pays depends on the answer being checkable against the
options, which a
      maths answer is and a GPQA/BBH answer is not)
Interpretation rule, frozen now: W1/W2 test whether the wave-6
volume jump
replicates on a second content set; if they fail, the causal
volume claim rests
```

```
on one dataset and must be reported as such. W3 tests the
two-factor reading
that emerged AFTER wave-6 and is therefore the genuinely
prospective part of
this round; a W3 miss falsifies the checkability explanation and
we revert to
reporting the direction of rewriting as unexplained.
At registration no job for these cells has been submitted and no
data exists.
squeue at registration:
    1730059     w5_06_q3_t7_c16384  R 1-12:03:02
    1735104          w6_q3_mc4_par  R    1:05:40
    1735106          w6_q3_mc4_seq  R    1:05:40
    1735714  w5_06_q3_t7_c16384_c2  R       1:53
    1729713        cs_m6_c16384_c2  R 1-13:32:45
```

ROUNDS 7–8 (ITERATIVE-FORCING CONDITIONS)

```
=== WAVE-8 ITERATIVE BUDGET FORCING, registered
2026-08-14T09:01:42Z (machine UTC) ===
Motivation: our sequential arm so far is a one-shot prompt suffix.
The studies we
analyse use ITERATIVE refinement / budget forcing. A reviewer can
object that we
criticise a protocol we never ran. This round implements iterative
continuation
forcing (generate; when the model tries to end its reasoning,
suppress the stop
and append "Wait"; continue until the per-sample budget is spent)
and asks
whether the papers conclusions survive it.
Setup: Qwen-7B and Qwen3-8B on AIME (60 q), N=8, per-sample budget
8192 tokens,
at most 4 forcings, matched against the SAME questions already
measured for the
one-shot suffix and for the natural parallel arm.
FROZEN PREDICTIONS:
  E1  Iterative forcing produces LONGER generations than the
one-shot suffix at
      the same budget: mean length ratio > 1.2, BOTH models.
  E2  Iterative forcing beats the one-shot suffix on majority-vote
accuracy at
      the same budget, BOTH models.
  E3  THE CENTRAL TEST. Iterative forcing at 8192 does NOT beat
the best natural
      parallel configuration on the same cell (the fair-baseline
comparison of
      the paper, which uses allowances up to 16384). I.e. the
reversal we report
      is not an artefact of using a one-shot suffix.
Interpretation rule, frozen now: E3 is the one that matters. If E3
holds, the
papers central claim extends to the operationalisation used in the
literature
and the limitation currently stated in the discussion can be
narrowed. If E3
fails -- forcing beats the fair parallel baseline -- then the
reversal IS
specific to the weak sequential arm, and we must say so in the
main text and
```

```
withdraw the general phrasing. E1/E2 are sanity checks that the
forcing
implementation does what it claims; if E1 fails the implementation
is suspect
and E3 carries no weight.
At registration no forcing job has been submitted and no forcing
data exists.
squeue at registration:


=== WAVE-9 ITERATIVE FORCING, REDESIGNED, registered
2026-08-15T13:26:44Z (machine UTC) ===
Wave-8 is recorded as 0/3 and is NOT reused. Its design was
faulty: at a total
budget of 8192 these models never emit an end-of-reasoning marker,
so there was
nothing for the forcing to suppress -- 60-73%% of samples were
never forced and
generated length was unchanged (ratio 0.99 / 1.01). E1 failed, so
by the rule
frozen in that round E3 carried no weight regardless of its value.
Redesign: total budget 16384, so the model can finish its
reasoning naturally
(it does so around 8-10k) and then be forced to continue, which is
what budget
forcing in the literature actually does. Everything else
unchanged: Qwen-7B and
Qwen3-8B, AIME 60 q, N=8, at most 4 forcings.
Reference points already measured on these same questions:
  one-shot suffix @16384   Qwen-7B 55.0%%   Qwen3-8B 71.7%%
  best natural parallel    Qwen-7B 61.7%%   Qwen3-8B 70.0%%
FROZEN PREDICTIONS:
  F1  the design now bites: mean forcings per sample >= 1.0, BOTH
models.
  F2  forcing generates longer than the one-shot suffix at the
same budget:
      length ratio > 1.15, BOTH models.
  F3  THE CORE TEST. Forcing does not exceed the best natural
parallel
      configuration by more than 5 points, for EITHER model.
Interpretation rule, frozen now: F1 is a precondition. If F1 fails
again the
round is void and we report in the paper that we could not
construct a forcing
condition that engages at these budgets -- we do not reinterpret
F3. If F1 holds
and F3 holds, the papers conclusion extends to iterative forcing
and the
limitation in the discussion narrows. If F1 holds and F3 fails,
the reversal we
report is specific to a one-shot suffix, and the main text must
say so.
At registration no wave-9 job has been submitted and no wave-9
data exists.
squeue at registration:
```