BEYOND THE FINAL ANSWER: CHARACTERIZING AND SELECTING REASONING TRAJECTORIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Sampling multiple reasoning trajectories creates two decisions: which unfinished trajectories deserve more computation, and which terminal answer should be submitted. These stages can benefit from different evidence. A trajectory can merit continuation before its current answer is reliable, while aggregate hidden-state geometry does not determine when substantial updates occur. We show that early answer-support histories predict later correctness even when both the current candidate and an isolated trial answer are wrong. Update timing also helps terminal comparison, including against aggregate-geometric controls. We introduce the Support History Score (SHS), combining candidate dominance frequency, confidence when leading, and historical competitive margin, and the Trajectory Outcome Score (TOS), based primarily on the temporal centroid of representational updates. SHS guides Continuation Selection (CS); two-stage Answer Selection (AS) retains two trajectories with SHS and compares their final submissions with TOS. Across three reasoning models and five mathematical and scientific benchmarks, support history improves continuation selection by 1.4–3.1 percentage points over matched endpoint-only scoring. Within the same retained pairs, TOS improves accuracy by 2.0–4.8 points over uniform choice. Both procedures reduce generated tokens while matching or exceeding majority-vote accuracy in most settings. The reasoning models remain unchanged; supervised two-input TOS extensions use problem-disjoint cross-fitting.

1 INTRODUCTION

Sampling multiple reasoning trajectories can improve LLM reasoning without updating model parameters (Wang et al., 2023; Karan & Du, 2026), but generation cost grows with both sample count and length, making computation allocation central to test-time scaling (Wu et al., 2025; Snell et al., 2025). This creates two sequential decisions: which unfinished trajectories deserve further computation, and which answer to submit after continuation. The first ranks trajectories by their potential for later success; the second compares final submissions.

Answer-confidence methods often assess whether a candidate exposed by the current trace is reliable. TRACE aggregates recent answer consistency and confidence to decide when to accept that candidate (Li et al., 2026). Continuation ranking instead concerns trajectories whose candidates may still be revised or replaced. We ask whether support history over complete candidate identities predicts later success, including when current-answer checks fail.

Latent trajectory methods expose a second distinction. *Tracing the Traces* (TTT) summarizes displacement, cumulative change, and directional alignment, and uses latent signals for learned early selection and sequential answer acceptance (Vilas et al., 2026). These summaries do not determine where large updates occur. Entropy Centroids preserves temporal location, but of high-entropy phases (Zhao et al., 2026). We ask whether the timing of representational updates informs terminal ranking beyond aggregate geometry.

Early support histories remain predictive when both the current top candidate and an isolated trial answer are wrong and, separately, when no historical frequency winner is correct. Update timing also helps terminal comparison, with benefits varying across models and retained pairs. We introduce the Support History Score (SHS) for continuation ranking and the Trajectory Outcome Score (TOS) for terminal comparison. Continuation Selection (CS) retains the highest-SHS trajectory; two-stage

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.

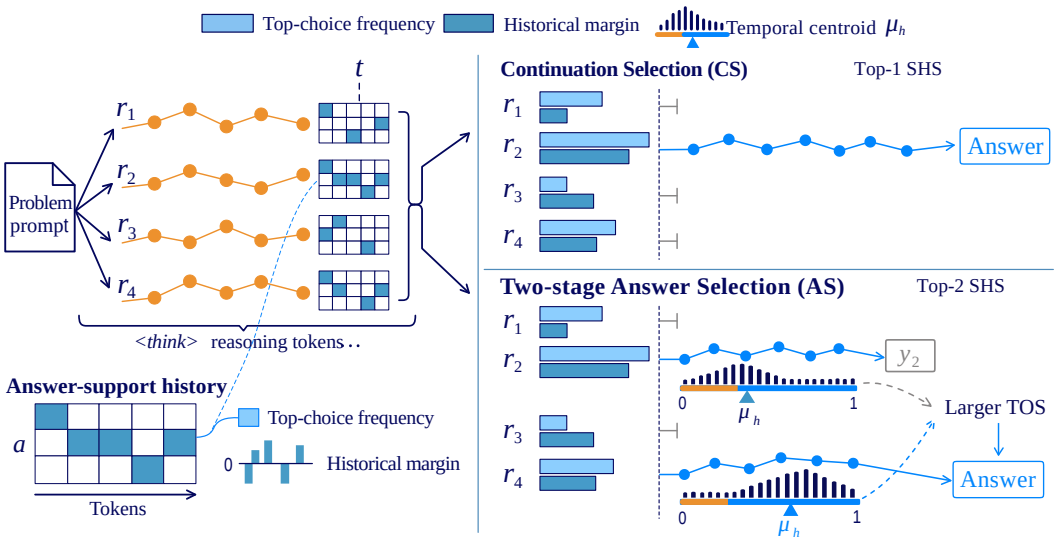


Figure 1: **Stage-specific trajectory selection.** SHS retains one trajectory for CS or two for AS. After continuation, AS submits a valid final answer using TOS. The illustrated signal components are defined in Section 3; the selection rules appear in Section 4.

Answer Selection (AS) retains two and selects between their final submissions using TOS. Both procedures reduce generated tokens while matching or exceeding majority-vote accuracy in most settings.

We make three contributions:

- **Distinguishing answer acceptance from continuation.** Early support histories rank later success even when current-answer checks fail; matched comparisons establish their benefit over endpoint-only scoring.
- **Preserving when representational change occurs.** Our temporal centroid captures information not determined by aggregate displacement, cumulative change, or directional alignment. We evaluate its selection value against geometry, length, and termination controls.
- **Stage-specific trajectory selection.** CS and two-stage AS use these signals for continuation and final selection. Matched controls separate early retention from terminal ranking and show how screening changes the value of terminal timing.

2 Related Work

Inference-time scaling and trajectory allocation. Inference-time scaling combines answer aggregation, search, and problem-dependent compute allocation (Wang et al., 2023; Wu et al., 2025; Snell et al., 2025). Learned verifiers assess solutions or intermediate steps (Cobbe et al., 2021; Lightman et al., 2024; Setlur et al., 2025). Adaptive self-consistency stops sampling using answer agreement (Aggarwal et al., 2023; Li et al., 2024); speculative rejection filters partial generations with reward models (Sun et al., 2024). ST-BoN uses cross-sample hidden-state consistency to retain one early trajectory and truncate the others (Wang et al., 2025a). Text-based behavioral analyses examine revision, redundant continuation, strategy switching, and how partial traces contribute to later reasoning (Lee et al., 2026; Chen et al., 2025; Wang et al., 2025b; Bachmann et al., 2026).

Answer confidence over reasoning. Answer reliability is estimated through model probabilities, verbalized confidence, semantic uncertainty, and confidence-weighted aggregation (Jiang et al., 2021; Xiong et al., 2024; Kuhn et al., 2023; Taubenfeld et al., 2025). Chain-of-Probe tracks confidence in the final predicted candidate across reasoning to rank completed traces (Wang et al., 2025c),

while TRACE aggregates recent answer consistency and confidence to decide when to accept a current answer (Li et al., 2026). SHS instead uses repeated complete-candidate readouts to rank unfinished trajectories by later outcomes. A trajectory may currently favor an incorrect candidate yet return an answer outside its earlier frequency-winning set.

Latent and temporal trajectory signals. Hidden-state probes extract correctness-related information from reasoning traces (Zhang et al., 2025). TTT summarizes displacement, cumulative change, and directional alignment, and applies latent-trajectory signals to learned early selection and sequential answer acceptance (Vilas et al., 2026). These aggregates do not retain the relative locations of update magnitudes along the observed trajectory. STEP and TrajSelector train hidden-state evaluators for trace pruning and final selection, respectively (Liang et al., 2026; Yu et al., 2025), while StALT constructs a training-free statistic from token-level hidden-state transitions (Furuya & Tanimura, 2026). Entropy Centroids weights the positions of high-entropy phases by their token lengths and favors lower centroids (Zhao et al., 2026). Our temporal centroid localizes representational updates for terminal comparison within the SHS-retained pair.

3 Measurable Differences Among Reasoning Trajectories

Early candidate support need not coincide with later trajectory success, and aggregate geometry does not determine when substantial updates occur. We study support histories before continuation and update timing afterward. For problem x, let r_t be the t-th reasoning token generated by model P_θ, with hidden representation $h_t^{(\ell)}$ at layer ℓ.

3.1 Representational Geometry and Update Timing

Displacement, cumulative update, and directional alignment summarize aggregate hidden-state geometry (Vilas et al., 2026). Reordering a fixed set of updates that share a common direction preserves these quantities while shifting substantial change toward the beginning or end. Our temporal centroid μ_h retains this timing information by weighting positions with update magnitudes, whereas Entropy Centroids localizes high-entropy phases (Zhao et al., 2026).

We divide the observed reasoning prefix into fixed-token blocks, retaining any shorter final block, and average the hidden representations within each block at each layer. This yields block states $z_1^{(\ell)}, \ldots, z_n^{(\ell)}$. Omitting the layer index, let the local updates be $v_i = z_{i+1} - z_i$ and the net displacement be $u = z_n - z_1$. We define

$$\widetilde{D}_h = \frac{\|u\|_2}{n}, \qquad U_h = \sum_{i=1}^{n-1} \|v_i\|_2.$$

Each statistic is computed per layer and averaged across valid layers. Appendix A.5 defines unnormalized displacement and directional alignment.

For $n \geq 2$, let c_i be the center of block i's actual token span, including any shorter final block, and let $d_{i,\ell} = \left\| z_{i+1}^{(\ell)} - z_i^{(\ell)} \right\|_2$.

We place each update midway between adjacent block centers and normalize its position over $[c_1, c_n]$. The trajectory score averages the magnitude-weighted temporal centroid across valid layers:

$$\xi_i = \frac{(c_i + c_{i+1})/2 - c_1}{c_n - c_1}, \qquad i = 1, \ldots, n-1,$$
$$\mu_h = \frac{1}{|\mathcal{L}_{\text{valid}}|} \sum_{\ell \in \mathcal{L}_{\text{valid}}} \frac{\sum_{i=1}^{n-1} d_{i,\ell} \xi_i}{\sum_{i=1}^{n-1} d_{i,\ell}}.$$

Here, $\mathcal{L}_{\text{valid}}$ contains layers whose cumulative update magnitude exceeds the numerical tolerance; μ_h is defined when this set is nonempty.

A higher μ_h indicates a later relative center of change (Figure 2). The score also depends on the observed endpoint: holding existing blocks and updates fixed, appending blocks with no additional change extends the normalization interval and reduces μ_h. The endpoint may be a natural end or an imposed observation boundary. Appendix A.5 gives the scale-invariance property and numerical conditions. Section 4.3 uses this centroid in the terminal score.

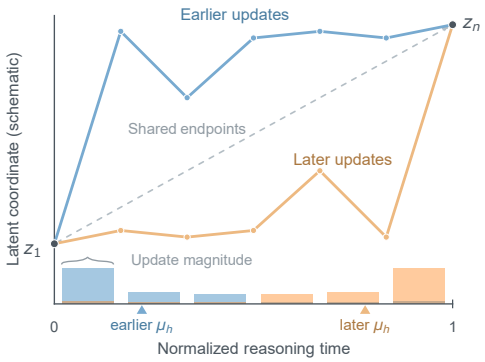


Figure 2: **Update timing.** Schematic hidden-state trajectories share endpoints but distribute update magnitudes differently over normalized reasoning time. Bars show hidden-state update magnitudes; triangles mark their temporal centroids.

3.2 Observing Answer Support

The natural next-token distribution $p_t(v) = P_\theta(v \mid x, r_{\leq t})$, with v ranging over the model vocabulary $\mathcal{V}$, predicts the continuation of the reasoning text. To observe answer support, we construct a standardized answer context R_t from the problem and observed reasoning prefix, and read the first-answer-token distribution $q_t(v) = P_\theta(v \mid R_t)$. These readouts use independent copies of the prefix, leaving the natural continuation unchanged.

To track complete answer identities across reasoning positions, we use a candidate-level distribution. For tasks with a well-defined finite answer set $\mathcal{A}$, let $\tau(a)$ denote the canonical sequence for candidate a, including its closing brace. We define

$$\rho_t(a) = P_\theta(\tau(a) \mid R_t), \qquad Z_t = \sum_{b \in \mathcal{A}} \rho_t(b), \qquad Q_t(a) = \frac{\rho_t(a)}{Z_t}.$$

Here, Z_t is the total probability mass assigned to the canonical answer set, and Q_t describes relative support within that set. A candidate favored by Q_t need not have been proposed as an answer to the full problem in the natural reasoning text.

For $P_t \in \{q_t, Q_t\}$, entropy $H(P_t) = -\sum_\omega P_t(\omega) \log P_t(\omega)$ summarizes support concentration but not answer identity. Complete-candidate readouts distinguish answers that share a first token; Appendix A.6 explains this distinction.

3.3 Answer-Support Dynamics

Two trajectories can have the same current readout top choice even if that candidate has repeatedly led in one history and only recently taken the lead in the other. We summarize these histories through top-choice frequency and historical margin.

Let $t_1 < \cdots < t_m \leq t$ be the recent measurement points available at observation position t. Splitting unit weight equally among tied top choices, we define

$$\mathcal{T}_j = \arg\max_{b \in \mathcal{A}} Q_{t_j}(b), \qquad w_j(a) = \frac{\mathbf{1}[a \in \mathcal{T}_j]}{|\mathcal{T}_j|},$$
$$f_t(a) = \frac{1}{m} \sum_{j=1}^{m} w_j(a).$$

Frequency records how often a candidate leads, but not the strength of its lead or the extent to which it trails at other points. We therefore define its historical margin against the strongest competing answer:

$$M_t(a) = \frac{1}{m} \sum_{j=1}^{m} \left[\log Q_{t_j}(a) - \max_{b \neq a} \log Q_{t_j}(b) \right].$$

The average covers all measurement points: contributions are positive when a strictly leads, zero when it shares the maximum, and negative when it trails. Thus, identical top-choice frequencies can conceal differences in both leading advantages and trailing deficits. Historical margin measures competition across the full window, rather than confidence only when a candidate is top-ranked. Both $f_t(a)$ and $M_t(a)$ are unchanged by reordering the observations within the window; they summarize history without encoding transition order.

The maximum top-choice frequency provides a trajectory-level summary, $s_f(t) = \max_{a \in \mathcal{A}} f_t(a)$.

The natural-trajectory case in Appendix A.7 illustrates identical top-choice sequences with different historical margins (Figure 7). These candidate-level quantities are components of the Support History Score (SHS), defined in Section 4.2.

SHS summarizes the support developed along a trajectory for continuation ranking, whose target is later success. Figure 3 shows its association with terminal outcomes; Section 5.3 evaluates the selection value of history against a matched endpoint-only score.

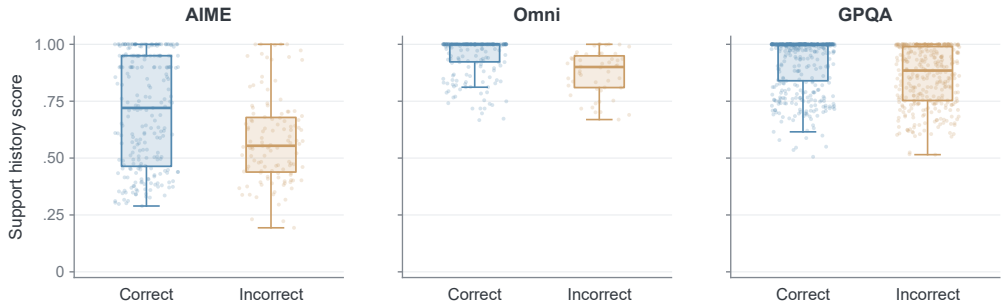


Figure 3: **Support history and trajectory outcomes.** Distributions of early SHS for correct and incorrect trajectories on AIME, Omni, and GPQA. The groups are more separated on AIME and overlap substantially on GPQA.

3.4 Current-Answer Reliability and Later Trajectory Success

Answer acceptance and continuation ranking evaluate different targets. A score may distinguish the correctness of an answer extracted from the current prefix more clearly than the outcome after additional reasoning.

We fix the low-entropy score $s_t = -H(q_t)$ and evaluate it against answers obtained from the observed prefix and after further reasoning. On Qwen3-4B trajectories from AIME 2024–2026, the score has higher within-problem AUC for current-answer correctness than for later correctness (Figure 4(a)).

We next evaluate SHS where both current-answer checks fail: the current top choice under Q_t is incorrect and an isolated trial answer from the same prefix is unsuccessful. Within this post hoc stratum, SHS remains predictive of later correctness (Figure 4(b)). In a separate stratum where no frequency-maximizing candidate is correct, SHS again ranks later outcomes; successful trajectories return an answer outside the earlier winner set.

A trajectory can therefore merit further computation without a trustworthy current answer. Appendix A.10 gives the evaluation protocols.

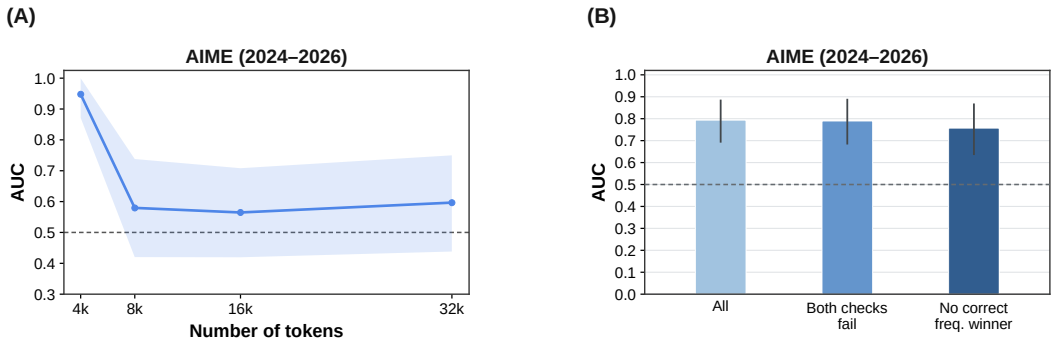


Figure 4: **Current answers and later success.** Within-problem AUC on Qwen3-4B, AIME 2024–2026. **(a)** A fixed early low-entropy score evaluated at current and later answer positions. **(b)** SHS predicts later correctness, including in two separately evaluated strata: both current-answer checks fail, or no frequency winner is correct. Shading and bars show 95% intervals; Appendix A.10 gives the protocol.

4 Inference-Time Selection Using Trajectory Signals

4.1 Selection Objectives and Shared Protocol

For each problem x, we independently sample K trajectories from P_θ with reasoning cap B: at a predefined observation position $t < B$, **Continuation Selection (CS)** retains one trajectory, whereas **Answer Selection (AS)** retains two for a final comparison. Retained trajectories continue until natural termination or B; Appendix A.2.2 specifies answer readout, tie handling, and token accounting. SHS ranks unfinished trajectories by their potential for later success; TOS compares final submissions within the retained set.

4.2 Support History Score for Continuation

The **Support History Score (SHS)** combines top-choice frequency, confidence when top-ranked, and historical margin. Frequency and confidence adapt TRACE's temporal stability signals (Li et al., 2026) from current-answer acceptance to ranking unfinished trajectories by later correctness. Historical margin adds candidate competition across the full window.

Window rule W selects the most recent measurement points ending at the observed prefix endpoint, using only available points for shorter histories. Using $f_t(a)$, $w_j(a)$, and $M_t(a)$ from Section 3.3, let $\mathcal{A}_t^+ = \{a \in \mathcal{A} : f_t(a) > 0\}$ be the candidates that have been top-ranked within the window. For these candidates, we compute the mean entropy-based confidence when top-ranked:

$$c_j = 1 - \frac{H(Q_{t_j})}{\log |\mathcal{A}|}, \qquad g_t(a) = \frac{\sum_{j=1}^m w_j(a) c_j}{\sum_{j=1}^m w_j(a)}, \quad a \in \mathcal{A}_t^+.$$

Thus $g_t(a)$ covers only top-ranked observations, whereas $M_t(a)$ covers the entire window.

For trajectory τ, we take the maximum candidate score:

$$\mathrm{SHS}_{\alpha,t}(\tau; W) = \max_{a \in \mathcal{A}_t^+} \left[\frac{1-\alpha}{2} \big(f_t(a) + g_t(a) \big) + \alpha \sigma(M_t(a)) \right]. \tag{1}$$

Here, $\sigma(u) = (1 + e^{-u})^{-1}$. We use two fixed recipes: $\alpha = 0$ gives weights $(1/2, 1/2, 0)$ on frequency, confidence, and margin, while $\alpha = 1/2$ gives $(1/4, 1/4, 1/2)$; these coefficients are not continuously optimized. When W contains only the final measurement, SHS reduces to a snapshot score. CS retains the highest-scoring trajectory; AS retains the two highest-scoring trajectories. Algorithm 2 in Appendix A.1 gives the CS procedure.

4.3 Trajectory Outcome Score for Final Selection

AS screens with $\mathrm{SHS}_{1/2}$, then compares valid submissions from the retained pair using the **Trajectory Outcome Score (TOS)**. The default score is raw μ_h, which retains the relative location of representational change. Three settings marked † in Table 1 combine it with cumulative update magnitude U_h through a two-input logistic extension:

$$\mathrm{TOS}^{(k)}(B) = \begin{cases} \mu_h^{(k)}(B), & \text{default,} \\ \sigma\Big(b + \beta_U z(U_h^{(k)}(B)) + \beta_\mu z(\mu_h^{(k)}(B))\Big), & \text{marked settings.} \end{cases} \tag{2}$$

The extension uses five problem-disjoint folds. For each model, logistic regression is fitted on four folds pooled across the five datasets; training-fold preprocessing and coefficients are frozen for the held-out fold. The reasoning model is unchanged. A uniquely valid answer is selected directly; otherwise TOS ranks the valid retained submissions. Appendix A.4 gives the development-selected configurations and fold-specific screening rules, and Appendix A.2.4 specifies fitting and score handling. Algorithm 1 gives the procedure.

5 Experiments

5.1 Experimental Setup

Models and tasks. We evaluate Qwen3-4B in thinking mode (Yang et al., 2025a), Phi-4-mini-reasoning (Xu et al., 2025), and DeepSeek-R1-Distill-Llama-8B (Guo et al., 2025) on AIME 2024–2026 (Mathematical Association of America, n.d.) (30 problems per year), Omni-64 (64 problems from Omni-MATH-Rule (Gao et al., 2025)), and GPQA-Diamond (Rein et al., 2024) (198 problems).

Generation and configuration. We independently sample $K = 4$ trajectories per problem and reuse these records across methods under a primary reasoning cap of $B = 32{,}768$ tokens. Appendix A.2 specifies generation settings, configuration selection, and evaluation folds.

Baselines. MV@4 selects the most frequent valid answer identity (Wang et al., 2023), while ST@4 selects the shortest reasoning trajectory after completing all four. We also compare with an early-selection adaptation of Chain-of-Probe (CoP; Wang et al., 2025c), LT-based early selection (LT-E), and the Combined variant of LT sequential acceptance (LT-C) (Vilas et al., 2026); see Appendix A.2.1.

Evaluation and accounting. Accuracy gives equal weight to each problem within a dataset, and generated tokens are reported in multiples of Single within each model–dataset pair. Evaluation and accounting details are given in Appendix A.2.

5.2 Main Results

Across the 15 model–dataset settings, CS and AS match or exceed MV@4 accuracy in **11** and **13** settings, respectively, while reducing generated tokens in every setting (Table 1). The largest CS gain is **10.0 percentage points** on Qwen3-4B AIME 2025. On Phi GPQA, LT-C achieves higher accuracy while generating fewer tokens than CS or AS.

5.3 Stage-Specific Evidence and Matched Controls

Support history beyond the endpoint. We replace each configured history with its final measurement, holding the observation endpoint, SHS weights, trajectory pool, and continuation protocol fixed. History improves CS for all three models by 1.42–3.13 percentage points (Figure 5, left). These gains arise from AIME; Omni-64 and GPQA already use a single measurement. Thus, history contributes beyond the matched endpoint score.

Separating early retention and terminal ranking. With TOS fixed, replacing random pair retention with SHS screening isolates early retention. Keeping the SHS-retained pairs fixed and replacing uniform selection over valid submissions with TOS isolates terminal ranking. Both interventions improve accuracy for Qwen3-4B and R1; Phi shows a terminal-ranking gain but only a small screening gain (Figure 5, middle and right).

Table 1: Accuracy (%) and generated tokens (×Single) under the 32K cap. CS/AS parentheses give accuracy gains (percentage points) and generated-token savings (%) over MV@4. **Bold** marks the highest accuracy and lowest generation per model–dataset pair, including ties. TOS uses raw μ_h by default; † marks the two-input logistic extension using (U_h, μ_h).

Method	AIME 2024		AIME 2025		AIME 2026		Omni-64		GPQA	
	Accuracy (%/Δpp)	Gen. tokens (×/saved%)	Accuracy (%/Δpp)	Gen. tokens (×/saved%)	Accuracy (%/Δpp)	Gen. tokens (×/saved%)	Accuracy (%/Δpp)	Gen. tokens (×/saved%)	Accuracy (%/Δpp)	Gen. tokens (×/saved%)
Qwen3-4B										
MV@4	76.67	4.00	66.67	4.00	63.33	4.00	84.38	4.00	56.31	4.00
ST@4	**80.00**	4.00	70.00	4.00	66.67	4.00	82.81	4.00	57.07	4.00
CoP	70.00	1.78	73.33	1.66	66.67	1.66	79.69	1.87	58.08	1.71
LT-E	73.33	**1.62**	66.67	**1.47**	63.33	**1.53**	**87.50**	**1.47**	56.06	**1.70**
LT-C	76.67	2.21	70.00	2.65	63.33	2.24	82.81	2.43	56.31	1.81
CS	**80.00** (+3.3)	1.70 (+57.6)	**76.67** (+10.0)	1.62 (+59.5)	**70.00** (+6.7)	1.71 (+57.3)	**87.50** (+3.1)	1.67 (+58.4)	59.60 (+3.3)	2.18 (+45.6)
AS	**80.00**† (+3.3)	2.45 (+38.8)	73.33 (+6.7)	2.33 (+41.8)	**70.00** (+6.7)	2.45 (+38.8)	84.38 (0.0)	2.38 (+40.5)	**61.11** (+4.8)	2.40 (+40.1)
Phi-4-mini-reasoning										
MV@4	45.00	4.00	40.00	4.00	41.67	4.00	71.88	4.00	53.28	4.00
ST@4	46.67	4.00	40.00	4.00	**46.67**	4.00	71.88	4.00	53.03	4.00
CoP	36.67	1.54	36.67	**1.39**	30.00	**1.45**	68.75	1.61	51.52	1.65
LT-E	36.67	1.60	33.33	1.60	26.67	1.61	71.09	1.87	48.48	1.91
LT-C	46.67	2.45	33.33	2.79	36.67	1.93	68.75	2.54	**57.58**	**1.45**
CS	**50.00** (+5.0)	**1.49** (+62.8)	36.67 (−3.3)	1.89 (+52.7)	43.33 (+1.7)	1.85 (+53.8)	70.31 (−1.6)	**1.60** (+59.9)	52.53 (−0.8)	2.16 (+46.1)
AS	46.67 (+1.7)	2.45 (+38.8)	**43.33** (+3.3)	2.30 (+42.6)	43.33 (+1.7)	2.40 (+39.9)	**75.00**† (+3.1)	2.62 (+34.4)	51.01† (−2.3)	2.49 (+37.8)
DeepSeek-R1-Distill-Llama-8B										
MV@4	50.00	4.00	35.00	4.00	43.33	4.00	69.53	4.00	47.73	4.00
ST@4	53.33	4.00	**43.33**	4.00	40.00	4.00	67.19	4.00	51.52	4.00
CoP	46.67	1.67	23.33	1.75	30.00	1.75	65.62	1.79	44.44	1.77
LT-E	53.33	**1.37**	23.33	1.42	**50.00**	**1.46**	63.28	1.82	49.49	**1.75**
LT-C	40.00	3.29	26.67	3.50	43.33	3.07	64.84	2.56	46.72	2.78
CS	**56.67** (+6.7)	1.86 (+53.5)	36.67 (+1.7)	**1.40** (+65.0)	43.33 (0.0)	1.57 (+60.8)	68.75 (−0.8)	**1.67** (+58.4)	49.49 (+1.8)	1.78 (+55.4)
AS	46.67 (−3.3)	2.22 (+44.5)	40.00 (+5.0)	2.45 (+38.7)	43.33 (0.0)	2.20 (+45.0)	**71.09** (+1.6)	2.22 (+44.5)	**52.27** (+4.5)	2.22 (+44.6)

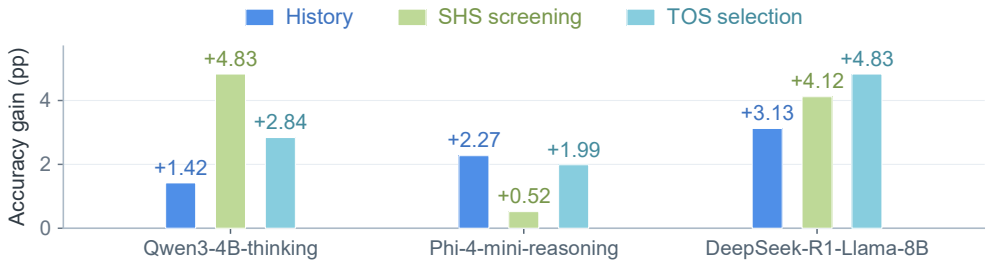


Figure 5: **Matched gains from history, screening, and terminal selection.** Bars show accuracy gains in percentage points for three matched comparisons: *History*, history-based versus endpoint-only SHS scoring in CS; *SHS screening*, SHS versus random pair retention with TOS fixed; *TOS selection*, TOS versus uniform choice among valid submissions within the same SHS-retained pairs. Accuracy gives equal weight to all 352 problems per model. The gains represent separate contrasts, not additive contributions.

Update timing and aggregate geometry. Within the same SHS-retained pairs, raw μ_h improves over uniform selection on all three models. Against the fixed geometric rule G_0, it is higher on Qwen3-4B and R1 but lower on Phi; full TOS exceeds G_0 in point estimate for all three. Timing can

thus inform terminal ranking, with selected settings benefiting from cumulative update magnitude as well. Length remains informative on Phi: the shorter-trajectory control reaches 54.26%, compared with 53.69% for TOS (Appendix A.3).

Conditional value of terminal timing. Relative to random pair retention, SHS screening increases the gain from pure μ_h selection on Qwen3-4B and R1, whereas Phi's interaction interval includes zero. The value of terminal timing therefore depends on the preceding allocation policy. Appendix A.3 reports the difference-in-differences estimates and intervals.

Signal discrimination. Early SHS distinguishes correct from incorrect outcomes more clearly than early geometric scores (Figure 6). Terminal geometric scores are more discriminative than their early counterparts, particularly on the mathematics tasks. Evaluation details and sensitivity analyses appear in Appendices A.10 and A.9.

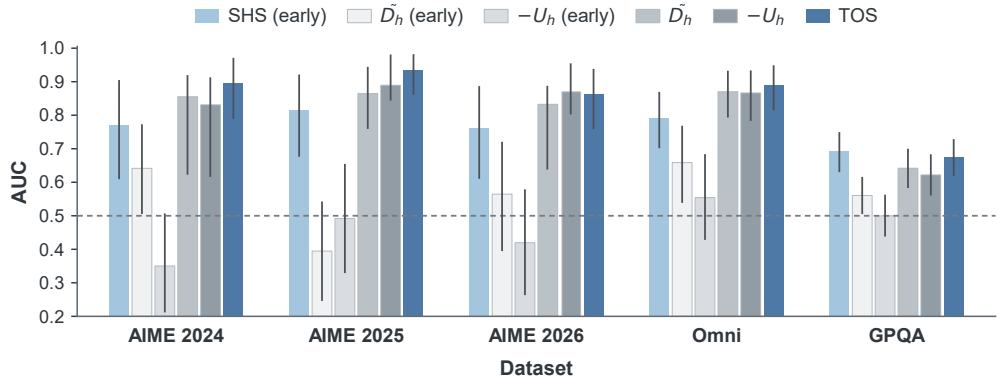


Figure 6: **Signal discrimination before and after continuation.** ROC-AUC for distinguishing correct from incorrect trajectory outcomes using answer-support and representational scores. Entries marked "early" use the observed prefix; other scores use the terminal trajectory. Higher values indicate better discrimination; the dashed line marks chance.

6 Discussion and Conclusion

Answer acceptance and continuation selection evaluate different targets. SHS remains predictive when both current-answer checks fail and, separately, when no historical frequency winner is correct. Histories may reflect developing results and plans, consistent with planning and backtracking in *Thought Anchors* (Bogdan et al., 2025). Retaining a trajectory need not entail accepting its present answer.

Aggregate geometry does not determine update timing. A higher μ_h may reflect late integration, revision, or a shorter low-change tail following substantial updates. Such timing may reflect learned tendencies in reasoning length, redundant continuation, and approach switching (Yang et al., 2025b; Chen et al., 2025; Wang et al., 2025b). Within SHS-retained pairs, raw μ_h improves over uniform selection on all three models. Its advantage over geometry and length controls varies; selected settings benefit from the two-input extension.

SHS screening increases the gain from pure μ_h selection on Qwen3-4B and R1; Phi's interaction interval includes zero. This motivates evaluating terminal signals conditional on prior allocation.

Limitations. SHS requires tractable canonical candidate sets, and TOS requires internal representations. Its two-input extension uses labeled outcomes with problem-disjoint cross-fitting. Reported savings measure generated tokens, excluding answer-support scoring and hidden-state extraction.

Our results support using support histories to allocate further computation and update timing to compare retained final submissions.

DECLARATION OF THE USE OF GENERATIVE AI (LARGE LANGUAGE MODELS)

Generative AI tools, including Grammarly and ChatGPT, were used solely for grammar checking and language polishing. All technical content, experimental design, data analysis, and conclusions were generated and verified exclusively by the human authors. The use of AI tools does not affect the originality or authorship of this work.

ETHICS STATEMENT

This work analyzes model-generated reasoning using public checkpoints and mathematical and scientific benchmarks; Appendix A.2 describes the Omni-64 subset. The methods aim to improve answer selection while reducing generated tokens. Selected outputs remain susceptible to errors and biases of the underlying models, and our evaluations characterize performance within the studied tasks and budgets.

REPRODUCIBILITY STATEMENT

Sections 3 and 4 define the signals and selection rules, with pseudocode in Appendix A.1. Section 5.1 identifies the models and benchmarks; Appendix A.2 documents budgets, measurement intervals, scoring configurations, and shared evaluation and token-accounting rules. Development selection uses a question set disjoint from evaluation. The two-input TOS extension uses five problem-disjoint folds; the appendix specifies preprocessing, fitting, and fold construction. Comparisons share trajectory records and answer protocols, including invalid-answer and exact-tie handling. Appendix A.10 specifies the AUC evaluation and problem-level bootstrap procedure.

REFERENCES

Pranjal Aggarwal, Aman Madaan, Yiming Yang, and Mausam. Let's sample step by step: Adaptive-consistency for efficient reasoning and coding with LLMs. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12375–12396, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.761. URL https://aclanthology.org/2023.emnlp-main.761/.

Gregor Bachmann, Yichen Jiang, Seyed-Mohsen Moosavi-Dezfooli, and Moin Nabi. The potential of CoT for reasoning: A closer look at trace dynamics. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/9c4c8d45f90ce37453db715dcda609fc-Abstract-Conference.html.

Paul C. Bogdan, Uzay Macar, Neel Nanda, and Arthur Conmy. Thought anchors: Which LLM reasoning steps matter? *arXiv preprint arXiv:2506.19143*, 2025. doi: 10.48550/arXiv.2506.19143. URL https://arxiv.org/abs/2506.19143.

Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. Do NOT think that much for 2+3=? On the overthinking of long reasoning models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 9487–9499. PMLR, 2025. URL https://proceedings.mlr.press/v267/chen25bx.html.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. doi: 10.48550/arXiv.2110.14168. URL https://arxiv.org/abs/2110.14168.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1): 1–26, 1979. doi: 10.1214/aos/1176344552. URL https://doi.org/10.1214/aos/1176344552.

Kotaro Furuya and Takahito Tanimura. Spatiotemporal hidden-state dynamics as a signature of internal reasoning in large language models. *arXiv preprint arXiv:2605.01853*, 2026. doi: 10.48550/arXiv.2605.01853. URL https://arxiv.org/abs/2605.01853.

Bofei Gao, Feifan Song, Zhe Yang, Zefan Cai, Yibo Miao, Qingxiu Dong, Lei Li, Chenghao Ma, Liang Chen, Runxin Xu, Zhengyang Tang, Benyou Wang, Daoguang Zan, Shanghaoran Quan, Ge Zhang, Lei Sha, Yichang Zhang, Xuancheng Ren, Tianyu Liu, and Baobao Chang. Omni-MATH: A universal olympiad level mathematic benchmark for large language models. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/f9e1e8b56c7e363985ebeb0e9dd1a85c-Abstract-Conference.html.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645:633–638, 2025. doi: 10.1038/s41586-025-09422-z. URL https://www.nature.com/articles/s41586-025-09422-z.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977, 2021. doi: 10.1162/tacl_a_00407. URL https://aclanthology.org/2021.tacl-1.57/.

Aayush Karan and Yilun Du. Reasoning with sampling: Your base model is smarter than you think. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/01498f17142547dadb60999c39bd0a09-Abstract-Conference.html.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=VD-AYtP0dve.

Seongyun Lee, Seungone Kim, Minju Seo, Yongrae Jo, Dongyoung Go, Hyeonbin Hwang, Jinho Park, Xiang Yue, Sean Welleck, Graham Neubig, Moontae Lee, and Minjoon Seo. The CoT encyclopedia: Analyzing, predicting, and controlling how a reasoning model will think. In *International Conference on Learning Representations*, volume 2026, pp. 141701–141743, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/e560202b6e779a82478edb46c6f8f4dd-Abstract-Conference.html.

Jiakun Li, Xingwei He, Kefan Li, Hongzheng Chai, Hongyue Yu, and Yuan Yuan. Efficient test-time scaling via temporal reasoning aggregation. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 13304–13318, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.651. URL https://aclanthology.org/2026.findings-acl.651/.

Yiwei Li, Peiwen Yuan, Shaoxiong Feng, Boyuan Pan, Xinglin Wang, Bin Sun, Heda Wang, and Kan Li. Escape sky-high cost: Early-stopping self-consistency for multi-step reasoning. In *International Conference on Learning Representations*, pp. 14751–14768, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/3fe2a777282299ecb4f9e7ebb531f0ab-Abstract-Conference.html.

Zhixiang Liang, Beichen Huang, Zheng Wang, and Minjia Zhang. Hidden states as early signals: Step-level trace evaluation and pruning for efficient test-time scaling. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 26800–26813, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.1336. URL https://aclanthology.org/2026.findings-acl.1336/.

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*, pp. 39578–39601, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/aca97732e30bcf1303bc22ac3924fd16-Abstract-Conference.html.

Mathematical Association of America. MAA Invitational Competitions: American Invitational Mathematics Examination (AIME). Official competition information, n.d. URL https://maa.org/maa-invitational-competitions/. Accessed 2026-09-26.

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 *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=Ti67584b98.

Amrith Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. Rewarding progress: Scaling automated process verifiers for LLM reasoning. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/98711dea460bdefe0e651ca23ec98ba2-Abstract-Conference.html.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/1b623663fd9b874366f3ce019fdfdd44-Abstract-Conference.html.

Hanshi Sun, Momin Haider, Ruiqi Zhang, Huitao Yang, Jiahao Qiu, Ming Yin, Mengdi Wang, Peter L. Bartlett, and Andrea Zanette. Fast best-of-N decoding via speculative rejection. In *Advances in Neural Information Processing Systems*, volume 37. Curran Associates, Inc., 2024. doi: 10.52202/079017-1025. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/3950f6bf5c2eb7435ecf58eaa85cc8c2-Abstract-Conference.html.

Amir Taubenfeld, Tom Sheffer, Eran Ofek, Amir Feder, Ariel Goldstein, Zorik Gekhman, and Gal Yona. Confidence improves self-consistency in LLMs. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 20090–20111, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.1030. URL https://aclanthology.org/2025.findings-acl.1030/.

Martina G. Vilas, Safoora Yousefi, Besmira Nushi, Eric Horvitz, and Vidhisha Balachandran. Tracing the traces: Latent temporal signals for efficient and accurate reasoning. In *International Conference on Learning Representations*, pp. 1776–1807, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/03e81b3bdb6c0a61093defd319d12203-Abstract-Conference.html.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc 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*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Yiming Wang, Pei Zhang, Siyuan Huang, Baosong Yang, Zhuosheng Zhang, Fei Huang, and Rui Wang. Sampling-efficient test-time scaling: Self-estimating the best-of-N sampling in early decoding. In *Advances in Neural Information Processing Systems*, volume 38, pp. 162137–162174. Curran Associates, Inc., 2025a. doi: 10.52202/085713-5415. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/ed45d6a03de84cc650cae0655f699356-Abstract-Conference.html.

Yue Wang, Qiuzhi Liu, Jiahao Xu, Tian Liang, Xingyu Chen, Zhiwei He, Linfeng Song, Dian Yu, Juntao Li, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. Thoughts are all over the place: On the underthinking of long reasoning models. In *Advances in Neural*

Information Processing Systems, volume 38, pp. 30591–30611. Curran Associates, Inc., 2025b. doi: 10.52202/085713-1027. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/2bfc4acb2e5c4af1576d14b0df9a664c-Abstract-Conference.html.

Zezhong Wang, Xingshan Zeng, Weiwen Liu, Yufei Wang, Liangyou Li, Yasheng Wang, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. Chain-of-Probe: Examining the necessity and accuracy of CoT step-by-step. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 2586–2606, Albuquerque, New Mexico, April 2025c. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-naacl.140. URL https://aclanthology.org/2025.findings-naacl.140/.

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*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/8c3caae2f725c8e2a55ecd600563d172-Abstract-Conference.html.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/6733cf15e10e2cd1d59af033c3bb8507-Abstract-Conference.html.

Haoran Xu, Baolin Peng, Hany Awadalla, Dongdong Chen, Yen-Chun Chen, Mei Gao, Young Jin Kim, Yunsheng Li, Liliang Ren, Yelong Shen, Shuohang Wang, Weijian Xu, Jianfeng Gao, and Weizhu Chen. Phi-4-Mini-Reasoning: Exploring the limits of small reasoning language models in math. *arXiv preprint arXiv:2504.21233*, 2025. doi: 10.48550/arXiv.2504.21233. URL https://arxiv.org/abs/2504.21233.

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. *arXiv preprint arXiv:2505.09388*, 2025a. doi: 10.48550/arXiv.2505.09388. URL https://arxiv.org/abs/2505.09388.

Shiming Yang, Yuxuan Tong, Xinyao Niu, Graham Neubig, and Xiang Yue. Demystifying long chain-of-thought reasoning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 71177–71209. PMLR, 2025b. URL https://proceedings.mlr.press/v267/yang25ae.html.

Bin Yu, Xinming Wang, Shijie Lian, Haotian Li, Changti Wu, Ruina Hu, Bailing Wang, Yuliang Wei, and Kai Chen. TrajSelector: Harnessing latent representations for efficient and effective best-of-N in large reasoning model. *arXiv preprint arXiv:2510.16449*, 2025. doi: 10.48550/arXiv.2510.16449. URL https://arxiv.org/abs/2510.16449.

Anqi Zhang, Yulin Chen, Jane Pan, Chen Zhao, Aurojit Panda, Jinyang Li, and He He. Reasoning models know when they're right: Probing hidden states for self-verification. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=O6I0Av7683.

Wenshuo Zhao, Qi Zhu, Xingshan Zeng, Fei Mi, Lifeng Shang, and Yi R. (May) Fung. Entropy centroids as intrinsic rewards for test-time scaling. *arXiv preprint arXiv:2604.26173*, 2026. doi: 10.48550/arXiv.2604.26173. URL https://arxiv.org/abs/2604.26173.

A Appendix

A.1 Selection Algorithms

Algorithm 1 Two-Stage Answer Selection with SHS and TOS

Input: Problem x, model P_θ, canonical candidates $\mathcal{A}$, $K = 4$, cap B, selected screening configuration (t, W), and frozen terminal rule R (raw centroid or fitted two-input logistic model)
Output: Final answer $\hat{a}$, or $\perp$ if no answer is validly formatted

Independently generate K trajectories until t or earlier natural endpoints.
Read canonical-answer probabilities from independent prefix copies using W; compute $\mathrm{SHS}^{(k)}_{1/2,t}$ for each trajectory.
Retain the two highest-scoring trajectories; exact boundary ties assign uniform probability over admissible retained pairs S_t.
Stop the others and continue unfinished trajectories in S_t until natural termination or B.
Obtain each $a^{(k)}_B$ using Appendix A.2.2.
If neither answer is valid, **return** $\perp$; **if** exactly one is valid, **return** that answer.
If the two valid answers are identical, assign uniform trajectory probability and **return** their common answer.
Compute the measurements required by R from all visible reasoning blocks; apply its frozen missing-value handling and, for logistic R, training-fold standardization and coefficients.
Select the larger TOS (Equation 2); exact ties assign uniform probability to both trajectories. **Return** the selected answer.
Evaluation integrates uniform choices by exact expectation.

Algorithm 2 Continuation Selection with Support History

Input: Problem x, model P_θ, canonical candidates $\mathcal{A}$, K trajectories, cap B, selected configuration (t, W, α) with $t < B$ and $\alpha \in \{0, 1/2\}$
Output: Selected trajectory's answer, or $\perp$ for an invalid answer

for $k = 1, \ldots, K$ **do**
    Generate naturally until t or an earlier reasoning endpoint.
    Use W to select available observations; read canonical-answer probabilities from independent prefix copies.
    Compute $\mathrm{SHS}^{(k)}_{\alpha,t}$ by Equation 1.
Assign uniform selection probability to trajectories tied for the largest SHS; let k^* denote a selected trajectory.
Stop further generation of all other trajectories.
if trajectory k^* is unfinished **then**
    Continue its original decoding process until natural termination or B.
Obtain $a^{(k^*)}_B$ using Appendix A.2.2.
return $a^{(k^*)}_B$ if validly formatted, otherwise $\perp$.
Evaluation integrates uniform choices by exact expectation.

A.2 Experimental Details

A.2.1 Task Construction and Baselines

Omni-64 is sampled from Omni-MATH-Rule problems (Gao et al., 2025) whose statements provably constrain the answer to an integer in 0–999, after excluding development problems and confirmed near-duplicates; the original problem statements are retained. The mathematical tasks use the integers 0–999 as answer candidates, whereas GPQA uses the four option labels. Results are reported separately for each dataset.

MV@4 selects the most frequent valid answer identity (Wang et al., 2023), and ST@4 selects the shortest reasoning trajectory after completing all four. We use an early-selection adaptation

of Chain-of-Probe (CoP; Wang et al., 2025c), LT-based early selection (LT-E), and the Combined variant of LT sequential acceptance (LT-C) (Vilas et al., 2026). LT-E uses a random forest over geometric trajectory summaries; LT-C processes trajectories in a fixed seed order and accepts an answer according to calibrated thresholds. LT fitting and calibration use problem-disjoint folds, and both LT and CoP are adapted to our observation and answer protocols.

A.2.2 Generation and Selection Protocol

We sample four independent trajectories per problem using each model's native reasoning template and the settings below. Natural generation uses BF16. All models use `min_p=0`, `repetition_penalty=1`, `presence_penalty=0`, and `frequency_penalty=0`, with a maximum context length of 40,960 tokens.

Parameter	Qwen3-4B	Phi-4-mini-reasoning	R1-Distill-Llama-8B
Temperature	0.6	0.8	0.6
Top-p	0.95	0.95	0.95
Top-k	20	Disabled (-1)	Disabled (-1)
Collection-request `max_tokens`	36,865	36,870	36,865

Primary and secondary evaluations use reasoning caps of 32,768 and 16,384 tokens, respectively. The collection-request limits reserve additional tokens for answer generation and boundary markers; they do not change the evaluated reasoning cap. Each trajectory uses an independently derived and recorded sampling seed. Trajectory indices 0–3 are identifiers, not sampling seeds. The 360 reused Qwen3-4B AIME trajectories use the same sampling settings and retain their original seeds.

Representations are averaged over actual 500-token blocks, including a shorter final block. Support observations are spaced by 500 reasoning tokens up to the selected boundary; an earlier natural endpoint is retained as the last observation. A window uses up to its stated maximum number of recent observations, with no padding after termination. Standard settings exclude the initial $t = 0$ probe. The two historical maximum-window-eight settings retain their original convention, which can include $t = 0$ for short histories. Thus a 1K boundary with maximum window three ordinarily contains only two positive-time observations.

At an observation, $\rho_j(a)$ is the unnormalized probability of the complete canonical answer string, including its closing delimiter, and $Q_j(a) = \rho_j(a)/\sum_{b\in\mathcal{A}} \rho_j(b)$. The entropy used by SHS is $H(Q_j) = -\sum_{a\in\mathcal{A}} Q_j(a) \log Q_j(a)$, over this normalized candidate distribution. Top-support ties split $w_j(a)$ equally. Window observations receive equal weight; confidence $g_t(a)$ is conditional on candidate a leading, while $M_t(a)$ includes every observation.

Continuation and answer readout. At the selected position $t < B$, CS retains one trajectory and AS retains two. Unfinished retained trajectories continue until natural termination or B. Let $a_B^{(k)}$ denote the answer from trajectory k under cap B. If reasoning ends naturally within the cap, the answer comes from the same sampled generation request. Otherwise, an isolated answer is generated from the prefix truncated at B, using greedy decoding with `temperature=0`, `top_p=1`, `top_k=-1`, and `max_tokens=256`. At the AS boundary, exact SHS ties are integrated uniformly over admissible retained pairs; CS uses uniform probability among tied highest scores. At termination, a uniquely valid retained answer is selected directly. With two distinct valid answers, the larger terminal score is selected. Identical answers or exactly equal scores receive uniform trajectory probability. If all retained answers are invalid, the submission is unsuccessful. Validity means accepted answer format, not correctness. There is no epsilon-based tie threshold: uniform decisions are evaluated by exact expectation, without a random tie-breaking seed.

When no trajectory ends early, CS and two-stage AS generate $Kt + (B - t)$ and $Kt + 2(B - t)$ reasoning tokens, respectively. Answer generation and signal measurement are accounted for separately.

A.2.3 Development Selection and Evaluation Folds

The development set contains 108 programmatically generated problems and 98 Omni-MATH-Rule problems outside Omni-64. It has no overlap with the evaluation problems. Development results determine the selected fixed or calibrated screening policy and terminal variant. Within calibrated policies, observation/window choices use training folds, and supervised terminal scores use held-out predictions.

We use five problem-disjoint folds, numbered 0–4, with seed 20260915. All four trajectories of a problem share a fold, and problem assignments are shared across models. Test-fold sizes are 71, 71, 71, 70, and 69. AIME retains its frozen historical assignment. Omni and GPQA are each ordered by SHA256 of the string `CS_AS-fold-v1|20260915|` followed by the problem ID, then assigned by index modulo five.

The standard observation/window menu is

$$\{(2000, 1), (2000, 3), (3000, 1), (3000, 3), (4000, 1), (4000, 3), (4000, 5)\}.$$

Additional settings include single-point and maximum-window-eight variants, and margin-inclusive configurations at 1K, 1.5K, and 2K. In CS, the three Qwen AIME years share a development-selected configuration. Qwen selection ties favor lower mean generated-token cost, then earlier observation, smaller window, and method identifier. Phi and R1 retain their development-selected, dataset-specific accuracy-maximizing settings, including historical representative choices among ties.

All AS screening uses $\alpha = 1/2$. TOP1_CAL selects the observation/window setting by training-fold four-to-one final accuracy; TOP2_CAL selects it by the training-fold probability that the retained pair contains at least one successful trajectory. Ties favor earlier observation, then smaller window. Calibration pools the three AIME years, treats Omni and GPQA separately, and is independent for each model. TOP1_CAL still retains two trajectories in AS: its name describes the calibration objective. R1 also uses fixed 1K/1.5K settings and previously evaluated expanded calibration settings. Appendix A.4 lists the selected rules.

A.2.4 Terminal-score Fitting and Preprocessing

Twelve main-table settings use the raw temporal centroid μ_h, without additional min–max or percentile normalization. Layers with cumulative update at most 10^{-10}, or no updates, are excluded. With exactly two blocks and a nonzero update, the centroid is $1/2$. A structurally missing trajectory centroid is replaced by the same-model, same-budget outer-training median, falling back to $1/2$ if no training value is available.

Qwen/AIME2024, Phi/Omni, and Phi/GPQA use logistic regression with exactly two inputs, (U_h, μ_h), as in Equation 2. There are no length, termination-status, displacement, or missingness-indicator inputs. Structural missing values are filled with training-fold medians; constant columns are removed. Retained features are standardized by the training-fold mean and population standard deviation. These preprocessing parameters are applied unchanged to the held-out fold. Both inputs are retained in all ten relevant fitted models.

For each model and held-out fold, fitting pools all five datasets from the other four folds, using all four trajectories per question with weight $1/4$ each; it is not restricted to retained pairs. Logistic regression uses L2 regularization with $C = 1$, the L-BFGS solver, at most 5000 iterations, tolerance 10^{-4}, and random state 20260915. No hyperparameter search is performed. Phi/Omni and Phi/GPQA share the same five fitted Phi models. The extension combines total update magnitude and temporal position.

Algebraic weight representation. The fitted coefficients satisfy $\beta_U < 0$ and $\beta_\mu > 0$ in all ten relevant models. Their within-fold ranking can therefore be written as

$$S_\lambda = (1 - \lambda)z(\mu_h) + \lambda z(-U_h), \qquad \lambda = \frac{-\beta_U}{\beta_\mu - \beta_U},$$

$$\operatorname{logit}(p_h) = b + (\beta_\mu - \beta_U)S_\lambda.$$

The values in Table 6 are algebraic summaries of fitted coefficients; no weight-grid search was performed. For raw-centroid settings, $\lambda = 0$ denotes the default rule. Because fold intercepts and scales differ, S_λ cannot replace the saved logistic probabilities in pooled-AUC analyses, even though the two give identical within-question rankings.

A.2.5 Evaluation and Accounting

Accuracy gives equal weight to each problem within a dataset; model-level accuracy averages over all 352 problems. Invalid answers count as unsuccessful. Policies replay the shared trajectory records and integrate uniform decisions by exact expectation, counting reasoning and answer tokens generated under each policy. The Single reference averages uniformly over individual trajectories. Table 1 reports generation in multiples of Single within each model–dataset pair; CS/AS parentheses give accuracy differences in percentage points and generated-token savings relative to MV@4. Signal-measurement costs are excluded and discussed in Appendix A.11.

A.3 Matched History and Terminal-Selection Controls

Matched history comparison. The history comparison in Figure 5 retains the same trajectory pool, observation endpoint, SHS component weights, and continuation and submission rules. Only the set of measurements in the window changes: SHS-last uses its final measurement, whereas SHS-history uses the configured window. Since the Omni-64 and GPQA settings already use one measurement, their results are identical under this intervention; the model-level improvements arise from the AIME settings. Accuracy gives equal weight to each of the 352 problems per model.

Terminal rules within the same retained pairs. Table 2 compares terminal rules after the same SHS screening. Uniform selection averages over the valid submissions in the retained pair. If both are valid, it assigns equal probability to each; if only one is valid, it selects that submission; if neither is valid, the problem is counted as incorrect. All other rules use the same valid-answer set. TOS uses raw μ_h in twelve settings and the model-level, five-fold logistic extension in the three marked settings, as described in Section 4.3. The shorter-trajectory control favors less generated reasoning, and the natural-termination control favors a trajectory that ended naturally over one stopped by the reasoning cap.

Fixed geometric control. The fixed geometric rule combines larger normalized displacement and smaller cumulative update with equal weights:

$$G_0(\tau) = \frac{1}{2}\, p_{\widetilde{D}}\Big(\widetilde{D}_h(\tau)\Big) + \frac{1}{2}\, p_{-U}(-U_h(\tau))\,.$$

Each component is mapped to its empirical percentile before averaging. For each held-out fold, the reference sample pools the other four folds across all five datasets for the same model and reasoning cap, including all four trajectories per problem. For a statistic $g \in \{\widetilde{D}_h, -U_h\}$, the mapping is

$$p_g(x) = \frac{1}{M_g} \sum_{j=1}^{M_g} \left[\mathbf{1}(g_j < x) + \frac{1}{2}\mathbf{1}(g_j = x) \right],$$

where M_g is the number of available training-trajectory values for that statistic. The mappings use no correctness labels and are applied unchanged to the held-out fold. Larger G_0 is preferred.

Table 2: Terminal-selection controls within the same SHS-retained pairs. Accuracy (%) gives equal weight to 352 problems per model. **Bold** marks the highest accuracy in each column, including ties.

Terminal rule	Qwen3-4B	Phi-4-mini	R1-8B
Uniform over valid submissions	65.91	51.70	48.58
Larger normalized displacement $\widetilde{D}_h$	66.61	52.27	51.56
Smaller cumulative update U_h	66.76	52.98	51.28
Shorter trajectory	67.05	**54.26**	51.99
Natural-termination preference	66.34	52.13	49.43
Higher temporal centroid μ_h	67.90	52.56	**53.41**
Fixed geometric rule G_0	67.61	53.27	51.70
TOS	**68.75**	53.69	**53.41**

The TOS differences from G_0 are $+1.14$, $+0.42$, and $+1.71$ percentage points for Qwen3-4B, Phi, and R1, respectively. Relative to the shorter-trajectory control, the differences are $+1.70$, -0.57, and $+1.42$ points for Qwen3-4B, Phi, and R1, respectively.

Interaction with the pure temporal centroid. Let $A(S, H)$ be accuracy in percent with pair-retention rule S and terminal rule H. The interaction analysis measures

$$\Delta_\mu = \big[A(\mathrm{SHS}, \mu_h) - A(\mathrm{SHS}, \mathrm{uniform})\big] - \big[A(\mathrm{random}, \mu_h) - A(\mathrm{random}, \mathrm{uniform})\big]. \quad (3)$$

Random retention averages uniformly over the six unordered pairs from the shared pool of four trajectories. The terminal comparison uses the pure temporal centroid for both retention rules.

Table 3: Interaction between SHS screening and pure μ_h terminal selection. Differences and 95% intervals are in percentage points.

Model	Δ_μ	95% interval
Qwen3-4B	$+2.53$	$[0.78, 4.29]$
Phi-4-mini-reasoning	-0.88	$[-2.68, 0.88]$
R1-Distill-Llama-8B	$+3.29$	$[1.02, 5.45]$

Random-pair control with TOS. The screening control in Figure 5 keeps the terminal selection configuration identical to AS and averages uniformly over all six unordered pairs from the four-trajectory pool. The uniform-final control retains the same SHS pair and changes only the terminal decision.

A.4 Selected Score Configurations

Table 4 gives all 15 CS configurations. Table 5 expands the AS observation/window choice for every held-out fold, including choices shared across folds. All AS settings use $\alpha = 1/2$. The three marked settings use the two-input logistic extension; the other twelve use the raw centroid. Table 6 reports the algebraic negative-magnitude weights, rounded to six decimals.

Table 4: Development-selected CS configurations. Window sizes are maxima; 1K denotes 1,000 reasoning tokens. The weights correspond to $(f, g, \sigma(M))$ in Equation 1.

Dataset	Observation	Max. window	Weights
Qwen3-4B			
AIME 2024	4K	5	$(1/4, 1/4, 1/2)$
AIME 2025	4K	5	$(1/4, 1/4, 1/2)$
AIME 2026	4K	5	$(1/4, 1/4, 1/2)$
Omni-64	3K	1	$(1/2, 1/2, 0)$
GPQA	4K	1	$(1/2, 1/2, 0)$
Phi-4-mini-reasoning			
AIME 2024	2K	3	$(1/4, 1/4, 1/2)$
AIME 2025	4K	8	$(1/4, 1/4, 1/2)$
AIME 2026	4K	3	$(1/4, 1/4, 1/2)$
Omni-64	2K	1	$(1/4, 1/4, 1/2)$
GPQA	4K	1	$(1/2, 1/2, 0)$
DeepSeek-R1-Distill-Llama-8B			
AIME 2024	4K	8	$(1/4, 1/4, 1/2)$
AIME 2025	2K	3	$(1/4, 1/4, 1/2)$
AIME 2026	3K	3	$(1/2, 1/2, 0)$
Omni-64	3K	1	$(1/2, 1/2, 0)$
GPQA	3K	1	$(1/4, 1/4, 1/2)$

Table 5: AS configurations by held-out fold. Each entry is the observation boundary/maximum window. TOP1_CAL and TOP2_CAL describe calibration objectives; AS always retains two trajectories. μ denotes raw-centroid selection, and LR† denotes logistic selection from (U_h, μ_h).

Dataset	Policy	Fold 0	Fold 1	Fold 2	Fold 3	Fold 4	TOS
Qwen3-4B							
AIME 2024	TOP2_CAL	4K/3	4K/3	4K/3	4K/3	2K/3	LR†
AIME 2025	TOP2_CAL	4K/3	4K/3	4K/3	4K/3	2K/3	μ
AIME 2026	TOP1_CAL	4K/5	4K/5	4K/5	4K/5	4K/5	μ
Omni-64	TOP1_CAL	3K/1	2K/1	2K/1	3K/1	3K/1	μ
GPQA	TOP2_CAL	2K/1	2K/1	2K/1	2K/1	2K/1	μ
Phi-4-mini-reasoning							
AIME 2024	TOP2_CAL	2K/3	2K/3	2K/3	2K/3	3K/3	μ
AIME 2025	TOP2_CAL	2K/3	2K/3	2K/3	2K/3	3K/3	μ
AIME 2026	TOP2_CAL	2K/3	2K/3	2K/3	2K/3	3K/3	μ
Omni-64	TOP2_CAL	4K/1	4K/1	4K/1	4K/1	4K/1	LR†
GPQA	TOP2_CAL	4K/3	4K/3	2K/1	2K/1	2K/1	LR†
DeepSeek-R1-Distill-Llama-8B							
AIME 2024	FIXED_1500	1.5K/3	1.5K/3	1.5K/3	1.5K/3	1.5K/3	μ
AIME 2025	TOP1_CAL	3K/3	4K/3	3K/3	3K/1	4K/5	μ
AIME 2026	FIXED_1000	1K/3	1K/3	1K/3	1K/3	1K/3	μ
Omni-64	FIXED_1000	1K/3	1K/3	1K/3	1K/3	1K/3	μ
GPQA	FIXED_1000	1K/3	1K/3	1K/3	1K/3	1K/3	μ

Table 6: Derived $\lambda = -\beta_U/(\beta_\mu - \beta_U)$ for the two-input logistic models. These are fitted-coefficient summaries, not searched hyperparameters. Phi's Omni-64 and GPQA settings share these models.

Fitted model	Fold 0	Fold 1	Fold 2	Fold 3	Fold 4
Qwen3-4B	0.641608	0.712073	0.676827	0.637822	0.719511
Phi-4-mini-reasoning	0.984137	0.905699	0.903955	0.979229	0.919118

A.5 Additional Geometric Statistics

Using the layerwise block states and updates from Section 3.1, the unnormalized endpoint displacement and net-direction alignment are

$$D_h = \|u\|_2, \qquad C_h = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \frac{v_i^\top u}{\|v_i\|_2 \|u\|_2},$$

where $\mathcal{I} = \{i : \|v_i\|_2 \|u\|_2 > 0\}$, and C_h is defined when $\mathcal{I} \neq \varnothing$. Each statistic is computed per layer and averaged across valid layers. The displacement scaled by block count is $\widetilde{D}_h = D_h/n$. Reordering updates that share a common direction preserves these aggregate quantities but can change the temporal centroid.

Centroid scale and numerical conditions. Because update magnitudes are normalized within each layer, multiplying all updates in a layer by the same positive factor leaves its centroid unchanged, provided the set of valid layers remains unchanged. Thus, μ_h measures the relative center of change independently of a uniform rescaling of its magnitude. It requires at least two blocks and at least one layer whose cumulative update exceeds the numerical tolerance 10^{-10}. With exactly two blocks and a nonzero update in a valid layer, the centroid is $1/2$. More generally, if only one update is nonzero, the layer's centroid equals that update's relative position ξ_i. Undefined trajectory scores use the same-model, same-budget training-fold median, with $1/2$ as the fallback when no training value is available.

A.6 Complete-Candidate and First-Token Readouts

Grouping candidates by their first token can hide changes in answer identity. Let $\phi(a)$ denote the first token of the canonical sequence $\tau(a)$, and define

$$Q_t^{\text{coarse}}(v) = \sum_{a:\phi(a)=v} Q_t(a).$$

This is a coarse-graining of Q_t, not the directly read distribution q_t. Two candidates sharing a first token can exchange their relative support while the probability of that token category remains unchanged. Tracking complete candidates preserves this distinction.

Canonical-answer scoring. We construct R_t from an independent copy of the original prompt and reasoning prefix through position t, appending the exact suffix

```
</think>\n\nThe final answer is \boxed{
```

This measurement leaves the natural continuation unchanged. The candidate set is the integers 0–999 for AIME and Omni-64, and the labels A/B/C/D for GPQA. The scored string includes the closing brace, for example `15}` or `A}`, but excludes EOS.

Let $\tau(a) = (s_1, \ldots, s_{L_a})$ be the token sequence of candidate a's canonical string. We compute

$$\ell_t(a) = \sum_{j=1}^{L_a} \log P_\theta(s_j \mid R_t, s_{<j}), \qquad Q_t(a) = \frac{\exp \ell_t(a)}{\sum_{b\in\mathcal{A}} \exp \ell_t(b)}.$$

Scores use the model's full-vocabulary softmax without temperature scaling, top-k or top-p filtering, or length normalization. The first-answer-token distribution is $q_t(v) = P_\theta(v \mid R_t)$, whereas Q_t is normalized over complete canonical answers. SHS confidence uses $H(Q_t)$.

The batched scorer uses FP32 forward computation, FP64 log-probability accumulation, and numerically stable normalization. It reuses the prefix KV cache, groups candidates by token length, and scores at most eight candidates per batch. Reused Qwen/AIME readouts use a trie-based implementation of the same scoring definition.

A.7 A Support-History Case

Figure 7 illustrates two natural trajectories with the same top-choice sequence: 5, 15, 15, 15, 20. Candidate 15 leads at three of five measurements in each trajectory, giving $f_t(15) = 3/5$. Its mean historical margins are 1.54 and -0.05, reflecting different leading advantages and trailing deficits despite identical frequencies. The trajectories later return 62 (correct) and 66 (incorrect), respectively. The example illustrates the information preserved by the margin definition; the matched history comparison in Section 5.3 evaluates the selection benefit across problems.

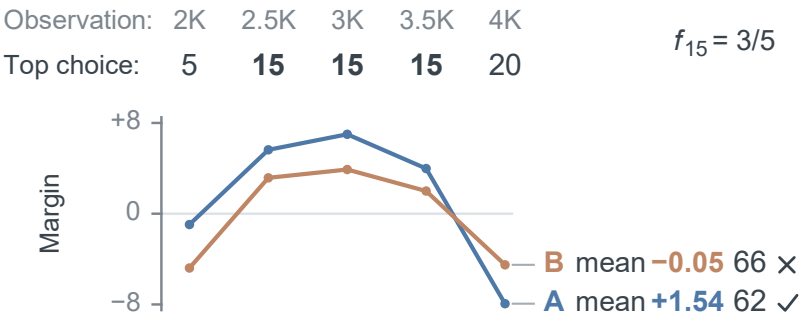


Figure 7: **Same frequency, different margins.** Candidate 15 leads at three of five points in both trajectories, with mean margins 1.54 and -0.05. Later answers are 62 (correct) and 66 (incorrect).

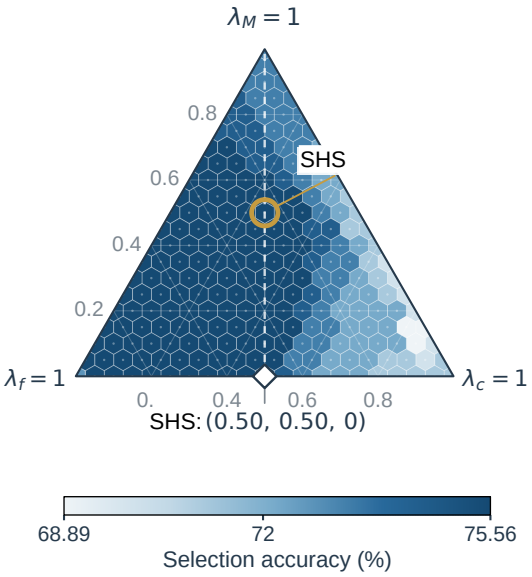
A.8 Score-distribution and Discrimination Protocols

Figure 3 groups early SHS on Qwen3-4B by terminal trajectory outcome. Figure 6 reports ROC-AUC by pooling trajectories within each dataset, comparing early answer support and geometry with terminal representational scores. The observation stage determines the prefix available for scoring; the prediction target is the terminal trajectory outcome.

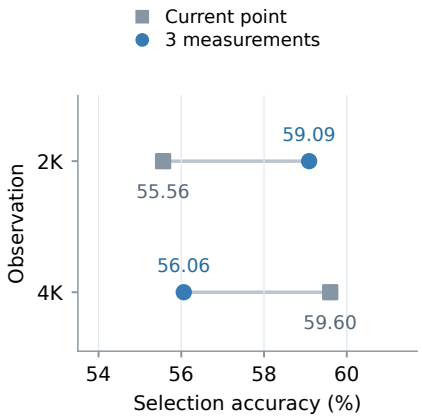
A.9 Sensitivity Analysis

Scoring weights. The weight-sensitivity comparison holds trajectories, observation positions, history windows, and the evaluation cap fixed while varying the relative weights of the three score components. Figure 8(a) shows a two-dimensional simplex grid of 231 fixed weight combinations evaluated on Qwen3-4B over the 90 AIME 2024–2026 problems, using four original trajectories per problem, a fixed 4K observation position, up to five actual history points, and the 32K reasoning cap. Selection accuracy ranges from 68.89% to 75.56%. SHS with weights $(0.50, 0.50, 0)$ and SHS with weights $(0.25, 0.25, 0.50)$ both attain $68/90$. Voronoi cell boundaries indicate display regions only; no performance interpolation is used.

Observation history. The benefit of support history depends on the observation position. On Qwen3-4B GPQA, with the observation position and margin-inclusive SHS weights fixed, using three measurements rather than the current measurement alone raises accuracy from **55.56%** to **59.09%** at 2,000 tokens. At 4,000 tokens, the ordering reverses: the current-point score reaches **59.60%**, compared with **56.06%** for three measurements (Figure 8(b)).



(a) Scoring weights



(b) Observation history

Figure 8: **Scoring weights and observation history. (a)** Selection accuracy across 231 fixed weight combinations on Qwen3-4B AIME 2024–2026. Both marked SHS configurations achieve $68/90$ correct selections. Cell boundaries delineate displayed weight regions. **(b)** Current-point versus three-measurement selection on Qwen3-4B GPQA at two fixed observation positions. Points show the reported accuracies.

A.10 Figure Evaluation Details

Figure 4 uses Qwen3-4B trajectories from the 90 AIME 2024–2026 problems, with four original trajectories per problem. We compute AUC separately for problems containing both successful and unsuccessful trajectories in the evaluated set, then average equally across these problems. Tied scores receive half credit. The 95% intervals use 10,000 whole-problem bootstrap resamples (Efron, 1979) stratified by year (seed 20260915).

In panel (a), the negative first-answer-token entropy is measured once at the early 4K prefix. Correctness is evaluated using the current isolated trial answer and answers under caps of 8,192, 16,384, and 32,768 tokens. Thus, the score is held fixed while the evaluation position changes. In panel (b), SHS with weights $(0.25, 0.25, 0.50)$ is measured at 4,000 tokens using up to five available history points, and later correctness is evaluated under the 32,768-token cap. The two conditional strata are evaluated separately: both the current candidate-level top choice and isolated trial answer are incorrect, or none of the frequency-maximizing candidates is correct.

A.11 Signal Measurement and Component Costs

Answer-support measurement teacher-forces canonical answer strings using batched candidate scoring and a shared prefix KV cache. Mathematical tasks have 1,000 candidate answers; GPQA has four. Hidden-state processing accumulates 500-token block representations and stores compact trajectory statistics without retaining all token-level activations.

The available offline timing records cover 3,864 additional trajectories. Their natural generation required approximately 50.23 GPU work-hours, compared with 1.93 GPU work-hours for geometric replay, a ratio of 3.84%. For GPQA, recorded complete-candidate scoring takes approximately 61–84 ms per measurement. These are component measurements from the offline pipeline; the generation-token results in Table 1 account for reasoning and answer generation. End-to-end latency including both signal measurements is not measured here.