SCM: DECISION-RELEVANT COMPARISONS FOR PROCESS REWARD MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

Process reward models (PRMs) guide reasoning by scoring intermediate steps and selecting among candidate solutions. Yet uniform pair weights ignore how often comparisons decide selection, while successful full-pool selection can conceal errors that become decisive when a stronger correct solution is absent. To train for these decisions, we study selection risk conditional on both correct and incorrect solutions being available. We propose Subset Champion Marginalization (SCM), an objective that weights comparisons by the joint distribution of the best available correct and incorrect solutions. With within-class orders fixed, these weights uniquely represent selection risk as linear comparison costs. SCM marginalizes these costs analytically across candidate budgets and combines the resulting trajectory loss with step supervision, using existing labels and the same inference-time scoring rule. We evaluate three reward-model backbones from 0.6B to 7B on four mathematical reasoning benchmarks through best-of-N selection and beam search. At 7B, SCM improves macro area under the log-budget accuracy curve on MATH-500 by 2.42 percentage points over the best matched baseline and 1.60 points over the best released PRM evaluated.

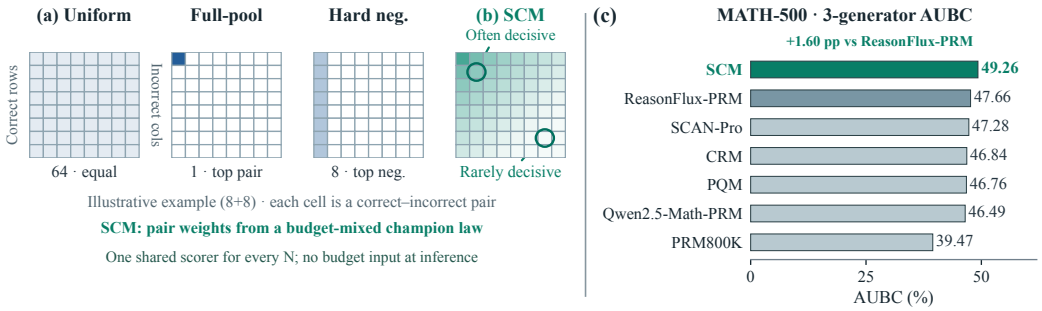


Figure 1: **SCM connects comparison weights to selection.** (a–b) Schematic weight maps: rows/columns index correct/incorrect trajectories in descending score order. Under the budget-mixed candidate law, each SCM weight is the joint probability that its correct–incorrect pair are the class champions; these weights uniquely reproduce selection risk for fixed within-class orders. The same scorer is used at every inference budget without a budget input. (c) MATH-500 macro area under the log-budget accuracy curve (AUBC) over three generators for all seven 7B PRMs in Table 1.

1 INTRODUCTION

Generally, scaling candidate generation yields reasoning gains only when models can reliably distinguish valid solutions from incorrect ones within large candidate pools (Cobbe et al., 2021; Snell et al., 2025). Process reward models (PRMs) address this challenge by evaluating intermediate reasoning steps to guide best-of-N (BoN) selection and beam search (Lightman et al., 2024; Snell et al., 2025), which requires aligning step-level judgments with selection decisions over complete or partial trajectories (sequences of reasoning steps). This raises a fundamental question: how should PRM training prioritize comparisons whose decision impact varies with the available candidates?

Nevertheless, uniform pair weights overlook how frequently each comparison determines selection. Furthermore, restricting supervision to the strongest correct–incorrect pair in the full pool can conceal ranking errors that become decisive when a stronger candidate is absent. To address these

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.

limitations, we introduce candidate availability as a complementary dimension of PRM training, so that comparisons receive weight according to their influence on selection.

Therefore, we formulate PRM training as minimizing selection risk when both correct and incorrect candidates are available. Under a specified, score-independent distribution over candidate collections, only the best available trajectory of each class determines the outcome. To this end, we propose an objective, named Subset Champion Marginalization (SCM), using their joint distribution.

We evaluate SCM on MATH-500, GSM-Plus, Gaokao2023, and OlympiadBench using BoN and beam search. Following prior work (Lightman et al., 2024; Li & Li, 2025), BoN tests whether the verifier's preferred candidate is correct. With MetaMath-Mistral candidates, 7B SCM raises MATH-500 BoN@64 from the best matched baseline's 50.8% to 54.8%, a gain of 4.0 percentage points. Figure 1 compares 7B methods by the area under the log-budget accuracy curve (AUBC).

Contributions. Our main contributions are as follows:

(1) We formulate PRM training in terms of candidate availability and derive SCM weights that uniquely represent conditional selection risk for fixed within-class orders.
(2) We bound selection error in terms of surrogate loss and availability changes, and derive closed-form weights for training across budgets without subset sampling.
(3) We demonstrate that SCM improves BoN and beam search over matched Process Q-value Model (PQM)/Conditional Reward Modeling (CRM) at 7B and 1.7B, and achieves the highest macro BoN scores among evaluated 7B PRMs on MATH-500 and GSM-Plus.

2 Candidate Availability and Hidden Ranking Errors

A successful full-pool decision need not imply reliable selection from its subsets. We analyze final candidate-level outputs from three generators and five PRMs on frozen 128-candidate MATH-500 pools. Uniform without-replacement (WOR) subsets of sizes $8, 16, 32, 64$ expose changing availability. All 500 questions per generator are included, with generator means weighted equally (Appendix B.3).

Figure 2 shows a nonzero darker component for every PRM: subset errors persist on questions answered correctly from the complete pool. The lighter component captures errors on full-pool-wrong questions. Together, they measure selection error when a correct candidate is available; SCM has the smallest total at all four budgets. Thus, a smaller hidden component alone need not imply better selection: success depends on both kinds of error.

Separating coverage from ranking. For a question q, let $\mathcal{C}_q$ contain K labeled trajectories, partitioned into correct and incorrect sets $\mathcal{P}_q, \mathcal{N}_q$ of sizes a, b. A scorer S_θ selects the largest score, with a fixed deterministic total order resolving ties. A candidate collection is *mixed*, denoted M, if both outcomes occur. To separate coverage from selection, define

$$R^{\mathrm{wor}}_{q,n} = \mathbb{P}(\text{selected outcome is wrong} \mid \mathsf{M}, q, n) \tag{1}$$

for a uniform size-n subset sampled without replacement, $2 \le n \le K$. With $D_{q,n} = \binom{K}{n} - \binom{a}{n} - \binom{b}{n}$, error for fully scored pools decomposes as

$$R^{\mathrm{wor}}_{q,n,\mathrm{all}} = \underbrace{\frac{\binom{b}{n}}{\binom{K}{n}}}_{\text{no correct candidate}} + \underbrace{\frac{D_{q,n}}{\binom{K}{n}}}_{\text{mixed probability}} R^{\mathrm{wor}}_{q,n}. \tag{2}$$

Invalid binomial coefficients are zero, and conditional risk requires $D_{q,n} > 0$. All-negative collections cannot succeed; all-positive collections always succeed. Conditioning on mixedness isolates decisions that depend on scoring.

Without cross-class ties, complete separation, $\min S^+ > \max S^-$, ensures correct selection whenever a positive is available. Conversely, an inverted pair fails when exposed alone. This binary criterion leaves open *which errors to prioritize before separation*. SCM assigns each comparison its probability of deciding selection.

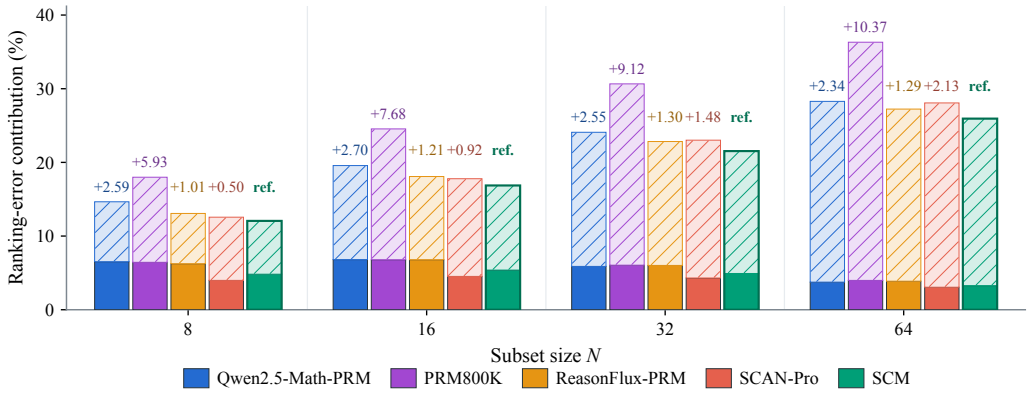


Figure 2: **Full-pool success hides subset errors.** All 500 questions per generator are included. Solid/hatched segments denote full-pool-correct/wrong questions; colors identify PRMs. Equal-generator means omit shared no-correct-candidate errors, preserving within-budget accuracy differences. Labels give SCM's expected-accuracy gains in percentage points (pp), computed from final candidate-level outputs. Details: Appendix B.3.

3 Subset Champion Marginalization

SCM trains for conditional selection risk. Under a specified, score-independent law over mixed candidate collections, class-champion probabilities weight trajectory comparisons, which complement step supervision (Figure 3).

3.1 Selection risk and the champion distribution

For valid step positions V_y with logits $z_{y,t}$, SCM uses the trajectory score and smooth gap penalty

$$p_{y,t} = \mathrm{sigmoid}(z_{y,t}), \qquad S_\theta(y) = \frac{1}{|V_y|}\sum_{t\in V_y} p_{y,t}, \qquad \phi_\tau(g) = \tau\log(1+e^{g/\tau}). \tag{3}$$

We use $\tau = 0.05$. Fix a labeled question group and a score-independent law Π_q over mixed candidate collections. In each collection A, the highest-scoring correct and incorrect trajectories are its two *class champions*. Their score gap, $G_q(A) = \max_{y\in A\cap\mathcal{N}_q} S_\theta(y) - \max_{y\in A\cap\mathcal{P}_q} S_\theta(y)$, determines selection. We therefore smooth this decision by $J_q(\theta;\Pi_q) = \mathbb{E}_{A\sim\Pi_q}\phi_\tau(G_q(A))$.

Order each class by score and the fixed tie rule as $p_1,\dots,p_a$ and $n_1,\dots,n_b$, with $S_i^+ = S_\theta(p_i)$ and $S_j^- = S_\theta(n_j)$. Let $I(A), J(A)$ be their champion ranks and e_{ij} indicate that n_j defeats p_i. The joint distribution $\mu_{ij} = \mathbb{P}_{\Pi_q}(I=i, J=j)$ supplies the weights $W_{ij} = \mu_{ij}$.

Theorem 1 (Champion sufficiency and exact risk representation)**.** *For any finite mixed-collection law, the nonnegative weights μ_{ij} sum to one and satisfy*

$$R_q(\Pi_q) = \sum_{i,j}\mu_{ij}e_{ij}, \qquad J_q(\theta;\Pi_q) = \sum_{i,j}\mu_{ij}\phi_\tau(S_j^- - S_i^+). \tag{4}$$

Indeed, the events $\{I=i, J=j\}$ partition mixed collections. Within each event, both the winner indicator and the champion gap are fixed, so total expectation yields the two identities. Thus candidate availability enters both objectives through the joint champion distribution: each comparison is weighted by its probability of deciding selection.

3.2 Selection-risk geometry and unique comparison costs

To characterize these comparison costs, fix the within-class orders and vary only their interleaving in the global ranking. Write $\mathcal{L}_{a,b}$ for these interleavings. Each $\sigma\in\mathcal{L}_{a,b}$ is uniquely specified by $k_i(\sigma)$, the number of negatives preceding p_i, where $0\le k_1\le\cdots\le k_a\le b$. Its inverted comparisons form

$$D_\sigma = \{(i,j) : j\le k_i(\sigma)\}, \qquad R_\mu(\sigma) = \sum_{(i,j)\in D_\sigma}\mu_{ij} = \mu(D_\sigma). \tag{5}$$

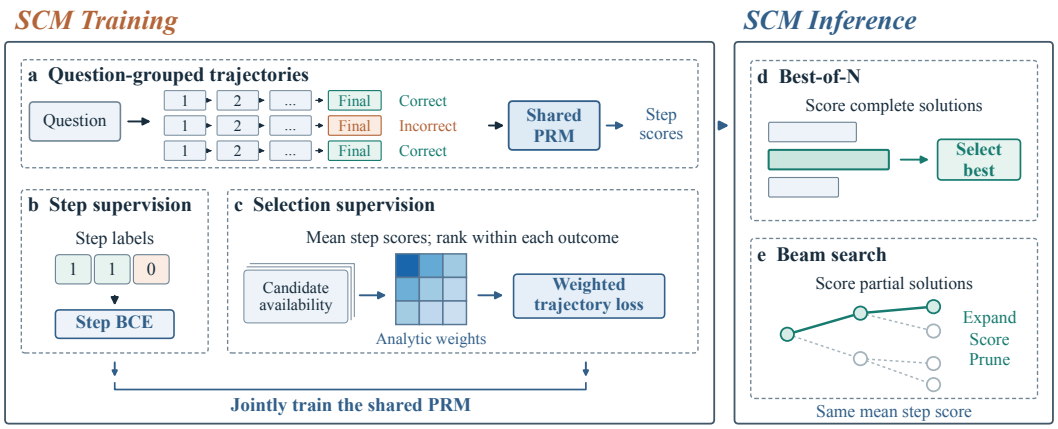


Figure 3: **SCM training and inference.** (a) A shared PRM scores steps within question-grouped trajectories. (b–c) Step binary cross-entropy (BCE) complements trajectory comparisons weighted by the probability that each correct–incorrect pair are the best available members of their classes. These probabilities are marginalized analytically across budgets; darker cells indicate greater weight. (d–e) The same mean step score ranks complete solutions for BoN and partial solutions for beam search. The weight map is schematic.

Since the class orders and collection law are fixed, μ is unchanged across these interleavings. Selection risk is therefore the probability mass of the inversion diagram.

Theorem 2 (Lattice structure and unique decision costs)**.** *Define $\sigma \vee \rho$ and $\sigma \wedge \rho$ by componentwise maximum and minimum of their k sequences. Then*

$$R_\mu(\sigma \vee \rho) + R_\mu(\sigma \wedge \rho) = R_\mu(\sigma) + R_\mu(\rho). \tag{6}$$

Correcting an adjacent inversion n_j, p_i reduces risk by exactly μ_{ij}. Moreover, if fixed coefficients V_{ij} satisfy $R_\mu(\sigma) = \sum_{i,j} V_{ij} e_{ij}(\sigma)$ for every $\sigma \in \mathcal{L}_{a,b}$, then $V_{ij} = \mu_{ij}$ for every pair.

The componentwise maximum and minimum remain nondecreasing, so both define valid interleavings. Their diagrams are respectively the union and intersection of the original diagrams; additivity of probability gives Equation 6. An adjacent cross-class swap changes only one inversion cell, giving the risk difference μ_{ij}. Finally, any chosen n_j, p_i can be made adjacent while preserving both class orders. Applying the proposed linear representation before and after that swap gives a difference of V_{ij}, which must equal μ_{ij}. Appendix A.1 gives the explicit construction.

Consequently, for two mixed laws with these fixed class orders,

$$R_\mu(\sigma) = R_\nu(\sigma) \text{ for every } \sigma \in \mathcal{L}_{a,b} \quad \Longleftrightarrow \quad \mu = \nu. \tag{7}$$

Equal risks imply equal adjacent-swap differences, recovering every probability; the converse follows from Equation 4. The champion distribution therefore retains exactly the information needed for this family of selection decisions.

This characterization distinguishes SCM from uniform and full-pool comparison rules. Uniform pairs assigns equal mass to all inversion cells; Full-pool pair places all mass on $(1, 1)$; Hardest negative distributes it along $j = 1$. SCM instead derives this mass from the candidate law. For $h = \mu_{11} < 1$, write $\mu = hE_{11} + (1 - h)T$, where E_{11} is the point mass at $(1, 1)$ and T is normalized elsewhere. When full-pool selection is correct, $e_{11} = 0$, and

$$R_\mu = (1 - h) \sum_{(i,j) \neq (1,1)} T_{ij} e_{ij} \leq 1 - h. \tag{8}$$

For $h = 1$, the risk is zero. Otherwise, hidden errors depend on the allocation of the remaining mass, which motivates the Matched top mass comparison in Section 4.4.

3.3 Surrogate control and changes in candidate availability

The representation above identifies the relevant comparisons. We next connect their smooth penalty to the probability of selecting an incorrect answer.

Theorem 3 (Excess-surrogate control of selection risk)**.** *For a fixed labeled pool and a score-independent mixed-collection law, independently assigned trajectory scores attain the minimum* $J_{\min} = \phi_\tau(-1)$ *on* $[0,1]$ *and have the same infimum on the sigmoid range* $(0,1)$*. Every score assignment satisfies*

$$R_q(\Pi_q) \le \frac{J_q(\theta;\Pi_q) - \phi_\tau(-1)}{\phi_\tau(0) - \phi_\tau(-1)}. \tag{9}$$

Every champion gap lies in $[-1,1]$, so $J_q \ge \phi_\tau(-1)$. Setting all positive scores to one and negative scores to zero attains equality on the closed domain; approaching these endpoints gives the same infimum on $(0,1)$. Whenever selection is wrong, including a tie resolved against the positive, its gap is nonnegative. Hence $\phi_\tau(G_q(A)) - \phi_\tau(-1) \ge [\phi_\tau(0) - \phi_\tau(-1)]\mathbf{1}[\text{wrong}]$; taking expectations proves Equation 9. As the surrogate approaches its free-score infimum, selection risk vanishes. Appendix A.2 also bounds errors with a prescribed negative-over-positive score margin.

Beyond a fixed law, the same representation quantifies changes in candidate availability. For two mixed laws Π_q, Π_q' on the same pool and fixed within-class orders, write $R_\mu(S), J_\mu(S)$ for Equation 4 under the induced distribution μ, and analogously for ν. Define

$$d_{\mathrm{sel}}(\mu,\nu) = \max_{\sigma \in \mathcal{L}_{a,b}} |\mu(D_\sigma) - \nu(D_\sigma)|. \tag{10}$$

This is the largest change in selection risk over the admissible interleavings. By Theorem 2, it vanishes precisely when $\mu = \nu$. In terms of total variation (TV),

$$\begin{aligned} d_{\mathrm{sel}}(\mu,\nu) &\le \mathrm{TV}(\mu,\nu) \le \mathrm{TV}(\Pi_q,\Pi_q'), \\ R_\nu(S) &\le \frac{J_\mu(S) - \phi_\tau(-1)}{\phi_\tau(0) - \phi_\tau(-1)} + d_{\mathrm{sel}}(\mu,\nu). \end{aligned} \tag{11}$$

The first line follows by restricting the events in total variation to inversion diagrams and then applying the champion map. The second combines the risk difference with Equation 9. It separates surrogate error under the chosen law from the effect of changing that law. Appendix A.3 proves the metric properties and gives an exact $O(ab)$ dynamic program for d_{sel}.

3.4 Closed-form marginalization across budgets

To compute these weights analytically, let $H_{ij} = \mathbb{P}(I \ge i, J \ge j)$ denote the joint tail. A rectangular difference recovers each pair's probability:

$$\mu_{ij} = H_{ij} - H_{i+1,j} - H_{i,j+1} + H_{i+1,j+1}. \tag{12}$$

The boundary tails vanish, so summation telescopes to $H_{11} = 1$. This identity reduces weight computation to counting tail events. Our training law draws $N \ge 2$ times uniformly with replacement (WR) from the K trajectories, conditioned on mixedness. It is defined even when $N > K$. With $r = K - i - j$,

$$W_{ij}^{\mathrm{wr}}(N) = \frac{(r+2)^N - 2(r+1)^N + r^N}{K^N - a^N - b^N}. \tag{13}$$

There are $K^N - a^N - b^N$ mixed draw sequences. Champions (i,j) require excluding higher-ranked members of each class and including both champions; inclusion–exclusion on the remaining $r+2$ candidates gives the numerator. Appendix A.4 derives the tail formula and its extension to unequal candidate probabilities.

SCM combines budgets through

$$W_{ij}^B = \sum_{N \in \mathcal{B}} \omega_N W_{ij}^{\mathrm{wr}}(N), \qquad \mathcal{B} = \{8,16,32,64,128\}, \quad \omega = \left(\tfrac{1}{8}, \tfrac{1}{4}, \tfrac{1}{4}, \tfrac{1}{4}, \tfrac{1}{8}\right). \tag{14}$$

Here ω is the normalized trapezoidal rule on the equally spaced $\log_2 N$ grid used by AUBC. More generally, any budget prior defines $\Pi_q^B = \sum_N \omega_N \Pi_{q,N}^{\mathrm{wr}}(\cdot \mid \mathsf{M})$ and, by linearity, $J_q^B = \sum_{i,j} W_{ij}^B \phi_\tau(S_j^- - S_i^+)$. The chosen outer integration matches AUBC; the training risk is mixed-conditioned, with coverage factors separated as in Equation 2. Resampling changes availability among the observed candidates.

Proposition 4 (From uniform comparisons to full-pool selection)**.** *For $a, b \geq 1$ and finite $N \geq 2$, all $W_{ij}^{\mathrm{wr}}(N) > 0$, $W_{ij}^{\mathrm{wr}}(2) = 1/(ab)$, and*

$$0 \leq 1 - W_{11}^{\mathrm{wr}}(N) \leq \min\left\{1, \frac{2(1-1/K)^N}{1-(a/K)^N-(b/K)^N}\right\} \longrightarrow 0. \tag{15}$$

At $N = 2$, every mixed pair has two draw orders. At finite budgets every pair can appear alone, while mass outside $(1, 1)$ requires at least one full-pool champion to be absent. A union bound on these two absence events gives Equation 15. Thus uniform pair weighting and full-pool supervision are endpoints of one availability family, with finite budgets retaining comparisons beyond the strongest pair. Appendix A.5 establishes convergence of both risk and surrogate.

3.5 Training objective and exact marginalization

Alongside the trajectory objective, SCM reuses RLHFlow Mistral/DeepSeek step labels: $b_{y,t} = 1$ if any of 16 sampled continuations from a prefix reaches a correct answer, and 0 otherwise (Appendix C.1). Binary cross-entropy fits each $p_{y,t}$ to its label, with $\mathcal{L}_{\mathrm{step}}$ averaged over valid labeled steps. The trajectory term then compares means from the same scorer. For mixed batch groups $\mathcal{Q}_{\mathsf{M}}$,

$$\mathcal{L}(W) = \mathcal{L}_{\mathrm{step}} + \lambda(t)\frac{1}{|\mathcal{Q}_{\mathsf{M}}|}\sum_{q\in\mathcal{Q}_{\mathsf{M}}}\sum_{i,j} W_{q,ij}\phi_\tau(S_j^- - S_i^+). \tag{16}$$

SCM uses $W = W^B$. The trajectory term is zero for empty $\mathcal{Q}_{\mathsf{M}}$; single-outcome groups still supply step BCE. Its coefficient $\lambda(t)$ ramps up after initial step supervision (Appendix C).

For optimization, W is constant within each region of fixed strict within-class orders. With $u_{ij} = \mathrm{sigmoid}((S_j^- - S_i^+)/\tau)$ and coordinate differences $d_{ij} = \mathbf{e}_j^- - \mathbf{e}_i^+$, the gradient is

$$\nabla_S J_q = \sum_{i,j} W_{ij}u_{ij}d_{ij} = \mathbb{E}_{A\sim\Pi_q}\nabla_S\phi_\tau(G_q(A)). \tag{17}$$

Each pass recomputes ranks with the fixed tie rule selecting a maximizing branch. Analytic weights avoid collection sampling; caching gives $O(K\log K + ab)$ cost per group beyond scoring, and Appendices A.8–A.9 detail curvature, variance bounds, and the algorithm.

4 Experiments

4.1 Experimental setup

Training and comparisons. We train SCM on Qwen2.5-Math-7B-Instruct, Qwen3-1.7B, and Qwen3-0.6B (Yang et al., 2024; 2025) with a common configuration: rank-64 low-rank adaptation (LoRA) (Hu et al., 2022) at 7B and full-parameter training for Qwen3. At 7B and 1.7B, PQM/CRM match SCM's data, adaptation regime, and training budget. Released comparators are: Qwen2.5-Math-PRM-7B, Qwen2.5-Math-7B-PRM800K, ReasonFlux-PRM-7B, and SCAN-Pro-7B.

Implementation. Training uses 200,000 trajectories in 12,428 question groups from RLHFlow Mistral/DeepSeek process data. SCM combines step BCE from continuation-success labels with terminal-label trajectory comparisons, averaging sigmoid step scores; baselines retain their objectives and aggregation. All trained models and ablations use final checkpoints (Appendix C).

Best-of-N evaluation. Llama3-70B-Instruct, MetaMath-Mistral, and Muggle-Math-13B each provide 128 frozen ordered candidates per question on MATH-500 (500 questions) and GSM-Plus (a 2,400-question subset) (Hendrycks et al., 2021; Lightman et al., 2024; Li et al., 2024b). At budget N, each scorer selects from the first N candidates with fixed labels, order, and tie-breaking. We summarize the five budgets by normalized area under the log-budget curve,

$$\mathrm{AUBC} = \frac{1}{4}\sum_{k=1}^{4}\frac{A_{2^{k+2}} + A_{2^{k+3}}}{2} = \frac{A_8 + 2A_{16} + 2A_{32} + 2A_{64} + A_{128}}{8}. \tag{18}$$

Macro AUBC averages the three generators equally, with differences calculated before rounding. Internal ablations use the same MATH-500 protocol (Appendix C.4).

Table 1: Best-of-N accuracy (%) on MATH-500 and GSM-Plus. PQM, CRM, and SCM match data, LoRA, and training budget, using their final checkpoints. Within each sampling policy, bold/underline mark the highest/second-highest distinct values per column across all seven 7B methods, including ties.

Sampling Policy	Methods	Dataset: MATH-500					Dataset: GSM-Plus				
		@8	@16	@32	@64	@128	@8	@16	@32	@64	@128
Llama3-70B-Instruct	Qwen2.5-Math-PRM-7B	52.4	53.0	53.6	**56.0**	56.4	74.58	75.54	75.46	75.62	75.58
	Qwen2.5-Math-7B-PRM800K	46.4	43.4	43.4	46.8	46.4	73.92	74.29	74.50	74.79	74.71
	ReasonFlux-PRM-7B	52.4	53.6	**54.8**	55.2	**58.0**	74.71	75.71	**75.92**	75.62	75.46
	SCAN-Pro 7B	52.0	53.2	51.0	50.2	52.4	74.50	74.88	75.63	75.38	75.54
	Qwen2.5-Math-7B-Instruct + PQM	51.4	52.6	51.6	53.4	54.2	74.58	75.71	75.63	75.71	75.63
	Qwen2.5-Math-7B-Instruct + CRM	51.6	52.6	52.8	53.6	53.2	75.04	75.63	75.63	75.92	**75.71**
	Qwen2.5-Math-7B-Instruct + SCM	**52.8**	**53.8**	**54.8**	55.6	56.8	**76.00**	**75.83**	75.71	**76.46**	**75.71**
MetaMath-Mistral	Qwen2.5-Math-PRM-7B	39.4	41.8	46.0	49.6	50.2	65.13	67.08	70.13	70.88	71.71
	Qwen2.5-Math-7B-PRM800K	37.2	40.0	42.4	45.6	43.0	62.79	65.71	65.92	67.42	68.08
	ReasonFlux-PRM-7B	41.2	44.0	48.2	51.4	51.6	66.25	**69.71**	**71.79**	**72.21**	72.83
	SCAN-Pro 7B	43.6	47.2	50.0	52.6	52.8	65.71	68.50	69.71	70.58	71.71
	Qwen2.5-Math-7B-Instruct + PQM	43.8	45.2	48.0	50.8	50.8	66.37	68.83	70.33	71.29	71.08
	Qwen2.5-Math-7B-Instruct + CRM	42.0	45.6	48.2	49.4	50.4	66.13	68.50	70.33	70.79	71.71
	Qwen2.5-Math-7B-Instruct + SCM	**44.6**	**48.2**	**51.8**	**54.8**	**54.8**	**67.13**	**69.71**	70.96	71.50	**72.88**
Muggle-Math-13B	Qwen2.5-Math-PRM-7B	33.4	37.0	39.4	43.2	44.8	63.83	66.88	68.42	70.83	70.92
	Qwen2.5-Math-7B-PRM800K	29.2	30.2	31.8	33.0	31.8	59.50	62.58	63.42	64.21	65.08
	ReasonFlux-PRM-7B	35.0	38.6	41.2	43.4	44.8	64.88	67.54	69.46	71.08	72.54
	SCAN-Pro 7B	35.2	38.2	40.8	43.8	44.8	64.54	67.29	68.83	69.71	70.00
	Qwen2.5-Math-7B-Instruct + PQM	35.0	39.2	39.2	42.2	42.6	64.79	67.58	68.67	70.63	71.29
	Qwen2.5-Math-7B-Instruct + CRM	35.6	**39.6**	40.0	43.0	41.8	64.96	**67.71**	68.96	70.83	70.96
	Qwen2.5-Math-7B-Instruct + SCM	**36.6**	39.4	**42.4**	**44.6**	**45.8**	**65.67**	67.63	**70.13**	**72.25**	**72.92**

Table 2: Best-of-N accuracy (%) with Qwen3 backbones. The 1.7B methods match data, full-parameter training, and budget, using final checkpoints. Bold/underline mark the top two distinct 1.7B values, including ties. The 0.6B row demonstrates applicability, with no matched baselines.

Sampling Policy	Methods	Dataset: MATH-500					Dataset: GSM-Plus				
		@8	@16	@32	@64	@128	@8	@16	@32	@64	@128
Llama3-70B-Instruct	Qwen3-1.7B + PQM	46.4	46.6	47.2	47.6	47.6	71.46	71.29	71.71	70.75	69.79
	Qwen3-1.7B + CRM	46.8	47.2	47.6	48.0	47.8	71.54	71.13	71.88	70.92	70.46
	Qwen3-1.7B + SCM	**47.4**	**47.8**	**49.2**	**48.2**	**48.8**	**71.88**	**71.79**	**72.33**	**71.58**	**70.75**
	Qwen3-0.6B + SCM	44.2	46.4	45.4	45.2	46.2	71.04	71.46	71.21	70.79	69.92
MetaMath-Mistral	Qwen3-1.7B + PQM	36.8	38.8	41.8	45.6	46.2	59.58	61.79	62.13	64.21	64.50
	Qwen3-1.7B + CRM	36.4	39.4	41.0	44.8	45.8	59.21	61.17	62.42	63.63	64.29
	Qwen3-1.7B + SCM	**37.4**	**40.6**	**43.6**	**46.8**	**47.8**	**60.83**	**62.79**	**63.75**	**65.33**	**65.54**
	Qwen3-0.6B + SCM	35.0	37.4	40.6	41.4	42.4	58.50	59.75	59.96	59.96	60.21
Muggle-Math-13B	Qwen3-1.7B + PQM	31.0	34.2	36.8	37.0	38.2	60.63	62.50	63.04	63.92	64.00
	Qwen3-1.7B + CRM	**31.4**	**34.8**	37.0	37.4	38.4	**61.04**	62.71	63.21	64.50	64.71
	Qwen3-1.7B + SCM	31.2	34.6	**37.6**	**38.4**	**39.6**	**61.04**	**62.92**	**64.08**	**65.62**	**65.54**
	Qwen3-0.6B + SCM	30.0	31.4	34.8	35.4	37.2	58.21	58.92	59.33	60.29	60.21

4.2 BEST-OF-N SELECTION

Consistent gains under matched training. At 7B (Table 1), SCM reaches macro AUBCs of 49.26% on MATH-500 and 72.11% on GSM-Plus, gaining 2.42/0.73 percentage points (pp) over CRM and 2.50/0.76 pp over PQM. Gains hold in every pool; on MATH-500, the CRM gaps are 1.90, 3.78, and 1.58 pp for Llama, MetaMath, and Muggle.

The same pattern holds at 1.7B (Table 2): SCM reaches 42.74%/66.50% macro AUBC on MATH-500/GSM-Plus, gaining 1.03/0.90 pp over CRM and 1.18/0.97 pp over PQM. Across both backbones and tasks, SCM leads in generator-averaged accuracy at every budget. It wins all 60 generator–budget comparisons against PQM and wins 54, ties two, and loses four against CRM (Appendix B).

At 7B, SCM also leads matched baselines under ordered prefix, WOR, and WR, with 2.42–2.54 pp macro AUBC gains over CRM (Appendix B.5).

Comparison with released PRMs. Among the seven evaluated 7B PRMs, SCM leads macro AUBC on both benchmarks, exceeding ReasonFlux-PRM, the strongest released PRM, by 1.60 and 0.25 pp. It leads every budget on MetaMath's MATH-500 pool and four of five on Muggle (Table 1).

Table 3: Beam-search accuracy (%) using the BoN checkpoints. Columns identify 3B/7B generators and root candidate counts N_{root}. Bold/underline mark the highest/second-highest distinct values within each verifier-size group, including ties. Search configuration: Appendix D.

Verifier Root candidates N_{root}	Qwen2.5-3B-Instruct				Qwen2.5-7B-Instruct			
	MATH-500		Gaokao2023		MATH-500		Gaokao2023	
	8	16	8	16	8	16	8	16
ReasonFlux-PRM-7B	**76.40**	**77.80**	**64.75**	**66.06**	85.40	87.00	**72.58**	**74.41**
SCAN-Pro 7B	73.00	74.20	62.40	63.71	82.20	83.80	70.76	72.32
Qwen2.5-Math-PRM-7B	75.60	<u>76.80</u>	62.66	63.97	<u>85.60</u>	<u>87.20</u>	72.06	73.89
Qwen2.5-Math-7B-PRM800K	73.20	74.60	60.57	61.88	84.20	85.80	71.28	72.85
Qwen2.5-Math-7B-Instruct + PQM	72.40	73.40	62.14	63.19	83.60	85.00	69.45	71.02
Qwen2.5-Math-7B-Instruct + CRM	73.40	74.80	61.88	62.92	83.40	85.00	71.02	72.85
Qwen2.5-Math-7B-Instruct + SCM	<u>76.20</u>	**77.80**	<u>63.19</u>	<u>64.75</u>	**85.80**	**87.60**	<u>72.32</u>	<u>74.15</u>
Qwen3-1.7B + PQM	69.80	70.80	<u>56.92</u>	<u>57.96</u>	77.40	78.80	65.27	66.84
Qwen3-1.7B + CRM	<u>70.40</u>	<u>71.60</u>	56.14	57.44	<u>77.60</u>	<u>79.20</u>	<u>66.58</u>	<u>68.15</u>
Qwen3-1.7B + SCM	**71.00**	**72.40**	**57.44**	**58.75**	**78.20**	**79.80**	**67.89**	**69.71**
Qwen3-0.6B + SCM	69.80	71.00	56.92	57.96	76.40	77.80	65.80	67.36

Applicability at 0.6B. Qwen3-0.6B SCM reaches 39.63% MATH-500 and 63.39% GSM-Plus macro AUBC. Its BoN and beam results demonstrate applicability at a smaller scale; matched comparisons remain at 7B and 1.7B.

Olympiad-level reasoning. We further test seven 7B verifiers on 674 English OlympiadBench problems (He et al., 2024), ranking eight shared Qwen3-4B rollouts per question (thinking disabled). SCM reaches 35.76%, gaining 1.19/1.78 pp over PQM/CRM and ranking second behind Qwen2.5-Math-PRM-7B (Figure 4). Protocol: Appendix C.5.

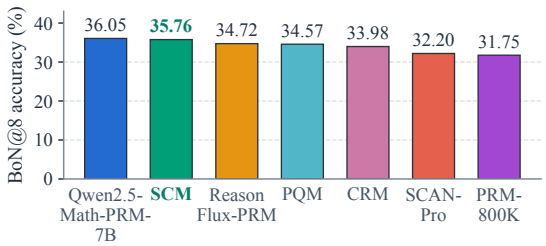


Figure 4: **OlympiadBench BoN@8 accuracy.**

4.3 PRM-GUIDED BEAM SEARCH

To test selection as partial solutions evolve, Table 3 uses the same BoN checkpoints with Qwen2.5-3B-Instruct and Qwen2.5-7B-Instruct generators on MATH-500 (500 questions) and Gaokao2023 (383 questions). All verifiers share the search settings: $N_{\text{root}} = 8/16$ first-layer candidates, beam widths $4/8$, two continuations per surviving beam, and at most 16 steps. Temperature is 0.7 with 2,048 tokens per step, allowing at most 128/256 action samples before early completion (Appendix D).

SCM 7B exceeds PQM, CRM, and Qwen2.5-Math-PRM-7B in all eight generator–task–budget cells. With the 7B generator, it reaches 85.80%/87.60% on MATH, gaining 2.40/2.60 pp over CRM. At 1.7B, SCM also surpasses PQM and CRM throughout, with the largest gains on Gaokao. Among released-model comparisons, SCM leads MATH with the 7B generator and ties ReasonFlux-PRM at $N_{\text{root}} = 16$ with the 3B generator; ReasonFlux-PRM leads all four Gaokao settings.

4.4 ABLATION STUDIES ON MATH-500

Table 4: Loss-component ablation. $\mathcal{L}^B_{\text{set}}$ is SCM's budget-weighted trajectory loss. MATH-500 AUBC (%); Macro averages generators equally. Bold/underline mark the top two distinct values.

Condition	$\mathcal{L}_{\text{step}}$	$\mathcal{L}^B_{\text{set}}$	Llama	MetaMath	Muggle	Macro
Step-only	✓		<u>52.60</u>	48.85	39.48	46.98
Set-only		✓	52.58	<u>49.33</u>	<u>41.35</u>	<u>47.75</u>
SCM	✓	✓	**54.75**	**51.13**	**41.90**	**49.26**

We examine supervision and comparison weighting through 7B rank-64 ablations on three frozen MATH-500 pools. Table 4 removes trajectory supervision (Step-only) or step BCE (Set-only, with its own ramp; Appendix C.2). Table 5 fixes both losses, score, temperature, and schedule, varying only W: the objectives in Section 3, without-replacement weights (Finite subsets), or reallocated non-champion mass (Matched top mass; Appendix A.6). Appendix B.2 gives full curves; Appendix B.4

Table 5: Comparison-weight objectives with both losses active: MATH-500 AUBC (%).

Condition	W_{ij}	Llama	MetaMath	Muggle	Macro
Uniform pairs	$(ab)^{-1}$	**55.10**	49.68	40.15	48.31
Full-pool pair	$(E_{11})_{ij}$	52.98	49.38	39.88	47.41
Hardest negative	$\mathbf{1}\{j=1\}/a$	53.10	49.58	40.08	47.58
Finite subsets	W^{U}_{ij}	54.38	<u>50.58</u>	<u>40.45</u>	<u>48.47</u>
Matched top mass	W^{C}_{ij}	54.00	49.95	40.40	48.12
SCM	W^{B}_{ij}	<u>54.75</u>	**51.13**	**41.90**	**49.26**

compares complete separation on 660 mixed validation groups, requiring every correct score to exceed every incorrect score.

Contributions of the two losses. SCM's 49.26% macro AUBC exceeds Step-only (46.98%) and Set-only (47.75%) by 2.28 and 1.51 pp. Its gains on every generator support combining step and trajectory supervision.

Which comparisons deserve weight? Uniform pairs reaches 48.31% macro AUBC, 0.95 pp below SCM, despite leading on Llama. Focusing on the strongest candidates also underperforms: Full-pool pair and Hardest negative reach 47.41% and 47.58%, trailing SCM by 1.85 and 1.68 pp. These results favor weighting comparisons by their selection probabilities.

Allocation beyond the champion pair. Finite subsets reaches 48.47%; fixing SCM's champion mass and using finite-subset residual weights yields 48.12% (Matched top mass). SCM's gains of 0.79/1.14 pp support weighting residual comparisons by selection probability (Equation 8).

5 Related Work

Process rewards for reasoning. Process supervision distinguishes intermediate correctness from final-answer success (Uesato et al., 2022; Lightman et al., 2024; Zhang et al., 2025). To scale supervision, prior work uses automated annotations (Wang et al., 2024; Luo et al., 2024; Lu et al., 2024) and progress targets (Setlur et al., 2025). PQM learns Q-value rankings (Li & Li, 2025), CRM models conditional failure times (Zhang et al., 2026), and ReasonFlux-PRM combines step and trajectory supervision (Zou et al., 2025). Self-denoising Monte Carlo annotation (SCAN) reduces noise (Ding et al., 2025); generative verifiers reason about step quality (Khalifa et al., 2026; She et al., 2025; Sohn et al., 2026), which ProcessBench evaluates through error localization (Zheng et al., 2025). SCM studies how candidate availability determines trajectory-level supervision.

Decision-focused learning and ranking. Decision-focused learning aligns predictions with downstream choices (Wilder et al., 2019). RankNet learns pairwise scores (Burges et al., 2005); ListNet, ListMLE, and SoftRank model ordered collections (Cao et al., 2007; Xia et al., 2008; Taylor et al., 2008), while LambdaLoss provides metric-driven objectives (Wang et al., 2018). Bipartite-ranking theory relates surrogate regret to the area under the receiver operating characteristic curve (AUC) (Agarwal, 2014). SCM uses candidate availability to derive unique conditional-risk weights for fixed within-class orders. Its analytic objective and excess-surrogate bound then link training to selection error.

6 Conclusion

In this paper, we introduce Subset Champion Marginalization (SCM), a process-reward objective that prioritizes comparisons by their role in candidate selection. By modeling candidate availability, SCM recovers the unique linear costs of conditional selection risk for fixed within-class orders. Analytic marginalization yields a weighted trajectory loss that complements step supervision. Experiments show improved BoN and beam search over matched PQM/CRM at 7B and 1.7B, with the highest macro BoN scores among evaluated 7B PRMs on MATH-500 and GSM-Plus. We hope this perspective motivates further research on process rewards that reflect candidate availability.

Reproducibility Statement

Section 3 specifies the score, availability law, objective, and core proofs; Appendix A gives supplementary derivations and the training algorithm. Appendices C and D record the training, BoN, and search configurations.

Ethics Statement

The study uses mathematical reasoning benchmarks and released automated annotations. Learned rewards do not guarantee mathematical correctness. Dataset and base-model use remains subject to the applicable source terms.

AI Use Statement

Artificial intelligence (AI) tools assisted with manuscript polishing, research ideation, and the generation of portions of the experimental code. The authors reviewed the AI-assisted text and code and take responsibility for the final manuscript, experiments, and conclusions.

References

Shivani Agarwal. Surrogate Regret Bounds for Bipartite Ranking via Strongly Proper Losses. *Journal of Machine Learning Research*, 15(49):1653–1674, 2014. URL https://jmlr.org/papers/v15/agarwal14b.html.

Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to Rank Using Gradient Descent. In *Proceedings of the 22nd International Conference on Machine Learning*, pp. 89–96, 2005. doi: 10.1145/1102351.1102363. URL https://doi.org/10.1145/1102351.1102363.

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to rank: From pairwise approach to listwise approach. In *Proceedings of the 24th International Conference on Machine Learning*, pp. 129–136, 2007. doi: 10.1145/1273496.1273513. URL https://doi.org/10.1145/1273496.1273513.

Jie Cheng, Gang Xiong, Ruixi Qiao, Lijun Li, Chao Guo, Junle Wang, Yisheng Lv, and Fei-Yue Wang. Stop Summation: Min-Form Credit Assignment Is All Process Reward Model Needs for Reasoning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 131646–131671, 2025. doi: 10.52202/085713-4382. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/be91eb86eb74efc055cff83e953f86ce-Abstract-Conference.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. URL https://arxiv.org/abs/2110.14168.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Yuchen Zhang, Jiacheng Chen, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, et al. Process Reinforcement through Implicit Rewards. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=9SkkifLopZ.

Yuyang Ding, Xinyu Shi, Juntao Li, Xiaobo Liang, Zhaopeng Tu, and Min Zhang. SCAN: Self-denoising monte carlo annotation for robust process reward learning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 54005–54033, 2025. doi: 10.52202/085713-1803. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/4dd0cfda282c778a21b38fd1d637f37b-Abstract-Conference.html.

Florian Eddie Dorner, Yatong Chen, André F. Cruz, and Fanny Yang. ROC-n-reroll: How verifier imperfection affects test-time scaling. In *International Conference on Learning Representations*, pp. 150895–150936, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/f4d943fad13060f97dee47e5200367a8-Abstract-Conference.html.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Xinyu Guan, Li Lyna Zhang, Yifei Liu, Ning Shang, Youran Sun, Yi Zhu, Fan Yang, and Mao Yang. rStar-Math: Small LLMs Can Master Math Reasoning with Self-Evolved Deep Thinking. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 20640–20661, 2025. URL https://proceedings.mlr.press/v267/guan25f.html.

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

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract-round2.html.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2106.09685.

Muhammad Khalifa, Rishabh Agarwal, Lajanugen Logeswaran, Jaekyeom Kim, Hao Peng, Moontae Lee, Honglak Lee, and Lu Wang. Process reward models that think. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=FPVCb0WMuN.

Maxim Khanov, Jirayu Burapacheep, and Yixuan Li. ARGS: Alignment as Reward-Guided Search. In *International Conference on Learning Representations*, pp. 17755–17773, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/4d1344317478ad99ff5f4e414aeab689-Abstract-Conference.html.

Peng Kuang, Yanli Wang, Xiaoyu Han, Yaowenqi Liu, Kaidi Xu, and Haohan Wang. Optimal aggregation of LLM and PRM signals for efficient test-time scaling. In *International Conference on Learning Representations*, pp. 92094–92109, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/94bd1b8c906494bada774096bd0fdd73-Abstract-Conference.html.

Xin Lai, Zhuotao Tian, Yukang Chen, Senqiao Yang, Xiangru Peng, and Jiaya Jia. Step-DPO: Step-wise Preference Optimization for Long-chain Reasoning of LLMs. *arXiv preprint arXiv:2406.18629*, 2024. URL https://arxiv.org/abs/2406.18629.

Chengpeng Li, Zheng Yuan, Hongyi Yuan, Guanting Dong, Keming Lu, Jiancan Wu, Chuanqi Tan, Xiang Wang, and Chang Zhou. MuggleMath: Assessing the Impact of Query and Response Augmentation on Math Reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10230–10258, 2024a. doi: 10.18653/v1/2024.acl-long.551. URL https://aclanthology.org/2024.acl-long.551/.

Qintong Li, Leyang Cui, Xueliang Zhao, Lingpeng Kong, and Wei Bi. GSM-plus: A comprehensive benchmark for evaluating the robustness of LLMs as mathematical problem solvers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2961–2984, 2024b. doi: 10.18653/v1/2024.acl-long.163. URL https://aclanthology.org/2024.acl-long.163/.

Wendi Li and Yixuan Li. Process reward model with Q-value rankings. In *International Conference on Learning Representations*, pp. 14708–14726, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/26494b66ae1114d314673e25b4967288-Abstract-Conference.html.

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.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1711.05101.

Jianqiao Lu, Zhiyang Dou, Hongru Wang, Zeyu Cao, Jianbo Dai, Yingjia Wan, Yunlong Feng, and Zhijiang Guo. AutoPSV: Automated Process-Supervised Verifier. In *Advances in Neural Information Processing Systems*, volume 37, pp. 79935–79962, 2024. doi: 10.52202/079017-2540. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/9246aa822579d9b29a140ecdac36ad60-Abstract-Conference.html.

Liangchen Luo, Yinxiao Liu, Rosanne Liu, Samrat Phatale, Meiqi Guo, Harsh Lara, Yunxuan Li, Lei Shu, Yun Zhu, Lei Meng, Jiao Sun, and Abhinav Rastogi. Improve Mathematical Reasoning in Language Models by Automated Process Supervision. *arXiv preprint arXiv:2406.06592*, 2024. URL https://arxiv.org/abs/2406.06592.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744, 2022. doi: 10.52202/068431-2011. URL https://papers.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.

Rattana Pukdee, Maria-Florina Balcan, and Pradeep Ravikumar. Reward learning from best-of-N preference data: Targets, tradeoffs, and design principles. *arXiv preprint arXiv:2605.30619*, 2026. URL https://arxiv.org/abs/2605.30619.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741, 2023. doi: 10.52202/075280-2338. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

Amin Rakhsha, Kanika Madan, Tianyu Zhang, Amir massoud Farahmand, and Amir Khasahmadi. Majority of the bests: Improving best-of-N via bootstrapping. In *Advances in Neural Information Processing Systems*, volume 38, pp. 37844–37877, 2025. doi: 10.52202/085713-1268. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/36556567e8437f137da23047309155dd-Abstract-Conference.html.

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*, pp. 60808–60838, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/98711dea460bdefe0e651ca23ec98ba2-Abstract-Conference.html.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Shuaijie She, Junxiao Liu, Yifeng Liu, Jiajun Chen, Xin Huang, and Shujian Huang. R-PRM: Reasoning-Driven Process Reward Modeling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 13438–13451, 2025. doi: 10.18653/v1/2025.emnlp-main.679. URL https://aclanthology.org/2025.emnlp-main.679/.

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*, pp. 10131–10165, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/1b623663fd9b874366f3ce019fdfdd44-Abstract-Conference.html.

Jiwoong Sohn, Tomasz Sternal, Kenneth Styppa, Torsten Hoefler, and Michael Moor. Process reward agents for steering knowledge-intensive reasoning. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2604.09482.

Mingyang Song, Zhaochen Su, Xiaoye Qu, Jiawei Zhou, and Yu Cheng. PRMBench: A fine-grained and challenging benchmark for process-level reward models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25299–25346, 2025. doi: 10.18653/v1/2025.acl-long.1230. URL https://aclanthology.org/2025.acl-long.1230/.

Mike Taylor, John Guiver, Stephen Robertson, and Tom Minka. SoftRank: Optimising Non-Smooth Rank Metrics. In *Proceedings of the International Conference on Web Search and Web Data Mining*, 2008. URL https://www.microsoft.com/en-us/research/publication/softrank-optimising-non-smooth-rank-metrics/.

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

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-Shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439, 2024. doi: 10.18653/v1/2024.acl-long.510. URL https://aclanthology.org/2024.acl-long.510/.

Xuanhui Wang, Cheng Li, Nadav Golbandi, Michael Bendersky, and Marc Najork. The LambdaLoss framework for ranking metric optimization. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pp. 1313–1322, 2018. doi: 10.1145/3269206.3271784. URL https://doi.org/10.1145/3269206.3271784.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *International Conference on Learning Representations*, 2023. URL https://research.google/pubs/self-consistency-improves-chain-of-thought-reasoning-in-language-models/.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022. doi: 10.52202/068431-1800. URL https://proceedings.neurips.cc/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

Bryan Wilder, Bistra Dilkina, and Milind Tambe. Melding the data-decisions pipeline: Decision-focused learning for combinatorial optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 1658–1665, 2019. doi: 10.1609/aaai.v33i01.33011658. URL https://ojs.aaai.org/index.php/AAAI/article/view/3982.

Fen Xia, Tie-Yan Liu, Jue Wang, Wensheng Zhang, and Hang Li. Listwise approach to learning to rank: Theory and algorithm. In *Proceedings of the 25th International Conference on Machine Learning*, pp. 1192–1199, 2008. doi: 10.1145/1390156.1390306. URL https://doi.org/10.1145/1390156.1390306.

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

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

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 11809–11822, 2023. doi: 10.52202/075280-0517. URL https://proceedings.neurips.cc/paper/2023/hash/271db9922b8d1f4dd7aaef84ed5ac703-Abstract.html.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T. Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. MetaMath: Bootstrap Your Own Mathematical Questions for Large Language Models. In *International Conference on Learning Representations*, pp. 45040–45061, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/c400474e8a36d0812fdee52739288b12-Abstract-Conference.html.

Zheng Zhang, Ziwei Shan, Kaitao Song, Yexin Li, and Kan Ren. Linking process to outcome: Conditional reward modeling for LLM reasoning. In *International Conference on Learning Representations*, pp. 87995–88016, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/8d6233829f97fe8eb3fa7d2ab774aedc-Abstract-Conference.html.

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

Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. ProcessBench: Identifying process errors in mathematical reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1009–1024, 2025. doi: 10.18653/v1/2025.acl-long.50. URL https://aclanthology.org/2025.acl-long.50/.

Jiaru Zou, Ling Yang, Jingwen Gu, Jiahao Qiu, Ke Shen, Jingrui He, and Mengdi Wang. ReasonFlux-PRM: Trajectory-aware PRMs for long chain-of-thought reasoning in LLMs. In *Advances in Neural Information Processing Systems*, volume 38, pp. 26764–26802, 2025. doi: 10.52202/085713-0901. URL https://papers.neurips.cc/paper_files/paper/2025/hash/26618fb384d3873b8ef6ab292a69095b-Abstract-Conference.html.

A Supplementary Theory and Training Algorithm

The main text develops the champion representation, its unique comparison costs, and the link between the smooth objective and conditional selection risk. We give the remaining constructions, availability formulas, and optimization details below. Throughout this appendix, candidate laws are score-independent and supported on mixed collections unless explicitly stated otherwise.

A.1 Interleavings and the uniqueness construction

Fix positive identities $p_1, \dots, p_a$ and negative identities $n_1, \dots, n_b$ in their respective within-class orders. Every total order preserving these two sequences corresponds to exactly one nondecreasing sequence $0 \le k_1 \le \cdots \le k_a \le b$. Conversely, place the first k_1 negatives before p_1, the next $k_2 - k_1$ before p_2, and so on, followed by the remaining negatives. This constructs the unique interleaving associated with k.

Componentwise minimum and maximum give the meet and join of this finite distributive lattice. Inversion diagrams preserve these operations as intersection and union. Equation 6 is consequently the modular identity for probability mass on the diagrams; it follows directly from the set identity for union and intersection.

For any pair (i, j), consider the interleaving

$$\sigma = (p_1, \dots, p_{i-1},\ n_1, \dots, n_{j-1},\ n_j, p_i,\ p_{i+1}, \dots, p_a,\ n_{j+1}, \dots, n_b). \tag{19}$$

Empty ranges are omitted. Swapping its adjacent n_j, p_i produces another admissible interleaving σ' and removes exactly the cell (i, j). Hence

$$\mu_{ij} = R_\mu(\sigma) - R_\mu(\sigma'). \tag{20}$$

If fixed coefficients V represent the risk for every admissible interleaving, the same subtraction recovers V_{ij}, proving $V = \mu$. This also proves a converse to Theorem 1: two mixed candidate laws yield identical selection risks for every interleaving if and only if their champion distributions agree. When they agree, their smooth objectives also agree for every score assignment preserving these within-class orders. The representation concerns conditional selection; the coverage terms in Equation 2 remain separate.

A.2 Excess-surrogate and margin bounds

For $\tau > 0$, ϕ_τ is strictly increasing. Since all score gaps belong to $[-1, 1]$, subtracting its minimum gives a nonnegative random variable. For $\gamma \in [0, 1]$ define $R_\gamma = \mathbb{P}_{\Pi_q}(G_q(A) \ge \gamma)$. Pointwise,

$$\phi_\tau(G_q(A)) - \phi_\tau(-1) \ge [\phi_\tau(\gamma) - \phi_\tau(-1)]\mathbf{1}[G_q(A) \ge \gamma], \tag{21}$$

and therefore

$$R_\gamma \le \frac{J_q - \phi_\tau(-1)}{\phi_\tau(\gamma) - \phi_\tau(-1)}. \tag{22}$$

Here γ measures the margin by which an incorrect champion exceeds the correct one. Selection error is bounded by R_0, with equality when every cross-class tie is resolved in favor of the negative. Equation 9 follows for any fixed tie rule.

Assigning all positives one and all negatives zero gives $G_q(A) = -1$ for every mixed collection. Thus the free-score minimum is $\phi_\tau(-1)$ and the minimum selection risk is zero. If a strict ordering or sigmoid range $(0, 1)$ is required, assign distinct positive scores approaching one and distinct negative scores approaching zero. Every champion gap then approaches -1, proving the same infimum. The resulting implication $J_q \to \phi_\tau(-1) \Rightarrow R_q \to 0$ is a finite-pool statement over independently assigned scores. The trained network optimizes the combined step and trajectory objective in Equation 16.

A.3 Equivalent laws and availability discrepancy

Let μ, ν be probability measures on the same $a \times b$ champion-rank grid, with fixed within-class identities and orders. Nonnegativity and symmetry of d_{sel} are immediate. For a third measure ξ, the triangle inequality follows from

$$|\mu(D_\sigma) - \nu(D_\sigma)| \le |\mu(D_\sigma) - \xi(D_\sigma)| + |\xi(D_\sigma) - \nu(D_\sigma)| \tag{23}$$

by maximizing over σ. If $d_{\mathrm{sel}} = 0$, every interleaving has the same risk under the two measures, and Equation 20 gives $\mu_{ij} = \nu_{ij}$ for each cell. Conversely, identical measures give zero discrepancy. Thus d_{sel} is a metric on champion distributions.

Total variation on this finite grid is

$$\mathrm{TV}(\mu, \nu) = \frac{1}{2} \sum_{i,j} |\mu_{ij} - \nu_{ij}| = \max_{E \subseteq [a] \times [b]} |\mu(E) - \nu(E)|. \tag{24}$$

Inversion diagrams form a subfamily of these events, proving $d_{\mathrm{sel}} \leq \mathrm{TV}(\mu, \nu)$. If the measures are induced by two mixed laws Π_q, Π'_q on a common collection space, let $T(A) = (I(A), J(A))$. Each grid event E corresponds to the collection event $T^{-1}(E)$, so

$$|\mu(E) - \nu(E)| = |\Pi_q(T^{-1}(E)) - \Pi'_q(T^{-1}(E))| \leq \mathrm{TV}(\Pi_q, \Pi'_q). \tag{25}$$

Repeated draws can be represented by their available identities, since multiplicity does not change either champion. This provides a common space of mixed presence sets for comparing with- and without-replacement laws.

For scores preserving the fixed class orders, Equation 5 gives $R_\nu(S) \leq R_\mu(S) + d_{\mathrm{sel}}(\mu, \nu)$. Substituting Equation 9 proves Equation 11. When class orders change, the champion maps and their induced measures are recomputed. The bound compares conditional selection risks on the same labeled support; changes in the probability of having a correct candidate are accounted for separately by coverage.

Exact computation. Let $\Delta_{ij} = \mu_{ij} - \nu_{ij}$ and $A_i(k) = \sum_{j=1}^{k} \Delta_{ij}$, with $A_i(0) = 0$. The largest signed risk difference equals

$$\max_{0 \leq k_1 \leq \cdots \leq k_a \leq b} \sum_{i=1}^{a} A_i(k_i). \tag{26}$$

Define $F_i(k)$ as the best sum through row i with $k_i = k$. Initialization and recurrence are

$$F_1(k) = A_1(k), \qquad F_i(k) = A_i(k) + \max_{0 \leq h \leq k} F_{i-1}(h), \quad 0 \leq k \leq b. \tag{27}$$

The constraint on the previous row is exactly $h \leq k$, so induction establishes the recurrence. Its final maximum is $M(\Delta) = \max_k F_a(k)$. Running it for $-\Delta$ as well gives $d_{\mathrm{sel}} = \max\{M(\Delta), M(-\Delta)\}$. Rowwise prefix maxima evaluate each recurrence in $O(b)$ time. The total cost is $O(ab)$ time and $O(b)$ working memory beyond the input.

A.4 Tail recovery and nonuniform candidate availability

For a general mixed law, write $H_{ij} = \sum_{k \geq i, \ell \geq j} \mu_{k\ell}$. Subtracting $H_{i+1,j}$ removes all rows except i; subtracting the analogous column tail then isolates μ_{ij}. This is the two-dimensional finite-difference form of Möbius inversion on the product of the two rank chains. Its boundary values $H_{a+1,j} = H_{i,b+1} = 0$ yield $\sum_{i,j} \mu_{ij} = H_{11} = 1$.

For uniform empirical sampling with replacement, set $a_i = a - i + 1$, $b_j = b - j + 1$, and $Z_N = K^N - a^N - b^N$. The relevant tails contain $a_i + b_j$ identities. Removing all-positive and all-negative sequences gives

$$H_{ij} = \frac{(a_i + b_j)^N - a_i^N - b_j^N}{Z_N}. \tag{28}$$

Applying Equation 12 recovers Equation 13. This calculation and the direct inclusion–exclusion argument count the same champion events.

The same argument allows unequal availability. Assign score-independent probabilities α_i to positive identity p_i and β_j to negative identity n_j, with class masses $p_+ = \sum_i \alpha_i > 0$, $p_- = \sum_j \beta_j > 0$, and $p_+ + p_- = 1$. Let

$$T_{ij} = \sum_{k=i}^{a} \alpha_k + \sum_{\ell=j}^{b} \beta_\ell. \tag{29}$$

For $N \geq 2$ independent draws conditioned on mixedness, the joint champion weights are

$$\mu_{ij}^{(N)} = \frac{T_{ij}^N - (T_{ij} - \alpha_i)^N - (T_{ij} - \beta_j)^N + (T_{ij} - \alpha_i - \beta_j)^N}{1 - p_+^N - p_-^N}. \tag{30}$$

All draws must fall in the two class tails, an event of probability T_{ij}^N. Excluding sequences missing p_i or n_j, then adding back those missing both, gives the numerator. The denominator is the mixed probability. Probabilities attach to candidate identities and follow them when ranks are recomputed; they do not depend on their numerical scores. Setting every identity probability to $1/K$ recovers the uniform formula exactly.

A.5 Finite-budget endpoints and objective convergence

Every pair has positive probability of being the only two identities drawn, with both appearing, for finite $N \geq 2$. This proves positive support. At $N = 2$, the numerator of Equation 13 is two and the denominator is $2ab$. For larger N, champions differ from $(1, 1)$ only if at least one of these two identities is absent. Each absence has unconditional probability $(1 - 1/K)^N$; the union bound divided by mixed probability proves Equation 15.

For fixed scores, set $\delta_N = 1 - W_{11}^{\mathrm{wr}}(N)$ and $g_{11} = S_1^- - S_1^+$. The risk and surrogate converge at the same residual-mass rate:

$$|R_{q,N}^{\mathrm{wr}} - e_{11}| \leq \delta_N, \qquad |J_{q,N}^{\mathrm{wr}} - \phi_\tau(g_{11})| \leq \delta_N. \tag{31}$$

Subtracting the $(1, 1)$ value in Equation 4 leaves only residual weights, which sum to δ_N. Indicator differences are at most one. For the surrogate, monotonicity and $\phi_\tau(1) - \phi_\tau(-1) = 1$ bound differences on $[-1, 1]$ by one as well. Positive support also implies that finite-budget risk is zero exactly when every positive precedes every negative in the selector's total order. Sampling the entire pool without replacement instead gives $W_{11} = 1$ immediately.

A.6 Comparison weights for internal controls

Table 4 includes Step-only, which removes the trajectory term, and Set-only, which retains SCM's set loss while omitting step BCE. All comparison-weight conditions in Table 5 retain both losses in Equation 16: Uniform pairs uses $W_{ij} = 1/(ab)$, Full-pool pair uses $W = E_{11}$, and Hardest negative uses $W_{ij} = \mathbf{1}[j = 1]/a$. These place, respectively, uniform mass on inversion cells, a point mass at $(1, 1)$, and uniform mass on the first negative column.

For a uniform subset of size n without replacement, both champions must be present and all higher-ranked candidates in each class absent. Choosing the other $n - 2$ trajectories below their respective champions gives

$$W_{ij}^{\mathrm{wor}}(n) = \frac{\binom{K-i-j}{n-2}}{D_{q,n}}, \qquad W_{ij}^U = \frac{1}{K-1} \sum_{n=2}^{K} W_{ij}^{\mathrm{wor}}(n). \tag{32}$$

The finite-subset reference W^U averages all feasible subset sizes. It changes both the availability law and the size prior relative to SCM. The Matched top mass condition holds the champion-pair weight fixed while changing its complement:

$$W^C = h^B E_{11} + (1 - h^B) T^U, \qquad h^B = W_{11}^B, \quad T_{ij}^U = \begin{cases} 0, & (i, j) = (1, 1), \\ W_{ij}^U / (1 - W_{11}^U), & \text{otherwise.} \end{cases} \tag{33}$$

For $a = b = 1$, the sole pair receives weight one. Otherwise $W_{11}^U < 1$ because the $n = 2$ component gives every pair positive mass. Under uniform sampling, h^B depends only on class counts and the budget prior, so the mass match holds for every ordering of the labeled group. Equation 8 explains the effect of reallocating the remainder.

A.7 Coverage and the order of budget conditioning

For empirical sampling with replacement, the mixed probability is $m_{q,N} = 1 - (a/K)^N - (b/K)^N$. The all-negative event has probability $(b/K)^N$, and all-positive collections contribute no error. Thus

$$\sum_{N \in \mathcal{B}} \omega_N R_{q,N,\mathrm{all}}^{\mathrm{wr}} = \sum_{N \in \mathcal{B}} \omega_N (b/K)^N + \sum_{N \in \mathcal{B}} \omega_N m_{q,N} R_{q,N}^{\mathrm{wr}}. \tag{34}$$

The scorer-dependent term includes the mixed probabilities $m_{q,N}$, whereas SCM's conditional risk is $\sum_N \omega_N R^{\mathrm{wr}}_{q,N}$. This is the with-replacement counterpart of Equation 2.

SCM first chooses a budget with probability ω_N and then draws a mixed collection at that budget. Choosing a budget, drawing an unconditioned collection, and then retaining mixed collections would instead produce the budget distribution

$$\widetilde{\omega}_{q,N} = \frac{\omega_N m_{q,N}}{\sum_{L\in\mathcal{B}} \omega_L m_{q,L}}. \tag{35}$$

Conditioning within each budget keeps the specified prior fixed across question groups with different label proportions.

A.8 Optimization geometry and exact marginal gradients

The softplus gap penalty belongs to the logistic pairwise-loss family used in preference-based reward fitting (Ouyang et al., 2022) and direct preference optimization (Rafailov et al., 2023). SCM's candidate law determines the comparison weights within this objective.

Fix a mixed group and a region of strict within-class score orders, where the champion weights are constant. With u_{ij} and d_{ij} as in Equation 17, differentiation gives

$$\nabla_S^2 J_q = \frac{1}{\tau}\sum_{i,j} W_{ij} u_{ij}(1-u_{ij}) d_{ij} d_{ij}^{\mathsf{T}} \succeq 0. \tag{36}$$

For any vector v, the quadratic form is $\tau^{-1}\sum_{i,j} W_{ij}u_{ij}(1-u_{ij})(v_j^- - v_i^+)^2$. All edge coefficients are positive for finite-budget SCM, so it vanishes exactly when every coordinate agrees. This bipartite graph Laplacian has nullspace $\mathrm{span}\{\mathbf{1}\}$, proving convexity within each ordering region and strict convexity there at fixed mean score.

Proposition 5 (Exact gradient marginalization)**.** *For a fixed mixed pool without within-class ties and a score-independent collection law, let $\widehat{g}_V$ average score-space gradients from V independent collections. Then*

$$\mathbb{E}[\widehat{g}_V] = \nabla_S J_q, \qquad \mathbb{E}\|\widehat{g}_V - \nabla_S J_q\|_2^2 \le \frac{2}{V}. \tag{37}$$

A collection with champions (i,j) has gradient $g(A) = u_{ij}d_{ij}$, with squared norm at most two. Partitioning by champions gives $\mathbb{E}g = \sum_{i,j} W_{ij}u_{ij}d_{ij}$. Independence makes cross terms between centered gradients vanish, yielding $\mathbb{E}\|\widehat{g}_V - \mathbb{E}g\|^2 = V^{-1}(\mathbb{E}\|g\|^2 - \|\mathbb{E}g\|^2) \le 2/V$. The analytic gradient removes this collection-sampling variance. At within-class ties the continuous objective can be nondifferentiable; the tie rule selects a maximizing branch. These statements concern score coordinates, which the neural model couples through its parameters.

A.9 Training algorithm

The model scores each trajectory once per training pass. Rank sorting and weight evaluation then operate on those scores. Backpropagation through the score gaps implements Equation 17, and the next pass recomputes ranks. For a fixed budget grid, weights can be cached by class counts; evaluating pair terms takes $O(ab)$ time and sorting takes $O(K\log K)$.

Algorithm 1 SCM for one optimizer batch

Require: Same-question groups $\mathcal{Q}$ with process and terminal labels
1: Score valid steps and trajectories using Equation 3.
2: Compute $\mathcal{L}_{\mathrm{step}}$ over valid labeled steps.
3: **for** each mixed group $q \in \mathcal{Q}$ **do**
4: Sort positive and negative trajectories by score with the fixed tie rule.
5: Obtain W_{ij}^B from Equations 13 and 14.
6: Compute $J_q = \sum_{i,j} W_{ij}^B \phi_\tau(S_j^- - S_i^+)$. ▷ Differentiate score gaps.
7: **end for**
8: Set $\mathcal{L}_{\mathrm{set}}$ to the average of J_q over mixed groups, or zero if none are present.
9: Update with $\mathcal{L}_{\mathrm{step}} + \lambda(t)\mathcal{L}_{\mathrm{set}}$.

B SUPPLEMENTARY RESULTS

The main text reports the complete BoN tables and AUBC comparisons. This appendix gives generator-averaged trends, full internal-ablation curves, and details of the selection diagnostics.

B.1 GENERATOR-AVERAGED BUDGET CURVES

Figure 5 visualizes the macro BoN curves reported in Tables 1 and 2. It makes the budget-dependent ordering visible without introducing an additional evaluation protocol. Multiple-sample reasoning also benefits from answer consensus (Wang et al., 2023; Shao et al., 2024). Verifier imperfection (Dorner et al., 2026), aggregation of generator and verifier signals (Kuang et al., 2026), and bootstrapped winner probabilities (Rakhsha et al., 2025) provide complementary accounts of how selection changes with inference compute; here, the curves compare trained scorers under fixed first-N selection.

B.2 COMPLETE INTERNAL-ABLATION CURVES

Tables 4 and 5 summarize the following eight conditions on Qwen2.5-Math-7B-Instruct with rank-64 LoRA. Step-only removes the trajectory term, Set-only removes step BCE, and the weighting controls retain both losses. All conditions use their final checkpoints after 1,554 optimizer steps (Appendix C). The common $N = 1$ result is verifier independent and is excluded from AUBC.

Table 6: Complete internal MATH-500 curves for Llama3-70B-Instruct. All values are accuracies (%) except the normalized AUBC summary. $N = 1$ is verifier independent and excluded from AUBC. Bold and underline indicate the two highest distinct values within each $N \geq 8$ column and AUBC.

Condition	@1	@8	@16	@32	@64	@128	AUBC
Step-only	37.4	51.8	52.6	53.0	52.6	52.6	52.60
Set-only	37.4	<u>53.2</u>	51.0	53.0	53.0	53.4	52.58
Uniform pairs	37.4	**53.6**	**54.0**	**55.2**	**56.4**	<u>56.0</u>	**55.10**
Full-pool pair	37.4	52.2	53.4	53.0	52.8	53.2	52.98
Hardest negative	37.4	52.0	53.0	53.4	53.6	52.8	53.10
Finite subsets	37.4	52.8	**54.0**	**55.2**	54.6	54.6	54.38
Matched top mass	37.4	52.2	<u>53.8</u>	54.2	54.4	55.0	54.00
SCM	37.4	52.8	<u>53.8</u>	<u>54.8</u>	<u>55.6</u>	**56.8**	<u>54.75</u>

Table 7: Complete internal MATH-500 curves for MetaMath-Mistral. All values are accuracies (%) except the normalized AUBC summary. $N = 1$ is verifier independent and excluded from AUBC. Bold and underline indicate the two highest distinct values within each $N \geq 8$ column and AUBC.

Condition	@1	@8	@16	@32	@64	@128	AUBC
Step-only	24.4	41.8	46.4	49.2	52.2	53.4	48.85
Set-only	24.4	43.0	**48.4**	50.0	51.4	52.0	49.33
Uniform pairs	24.4	42.6	47.2	50.2	53.0	54.0	49.68
Full-pool pair	24.4	42.4	47.2	50.2	52.4	53.0	49.38
Hardest negative	24.4	43.0	47.0	50.4	53.0	52.8	49.58
Finite subsets	24.4	<u>43.2</u>	47.8	**51.8**	<u>53.2</u>	**55.8**	<u>50.58</u>
Matched top mass	24.4	42.8	47.4	<u>50.8</u>	<u>53.2</u>	54.0	49.95
SCM	24.4	**44.6**	<u>48.2</u>	**51.8**	**54.8**	<u>54.8</u>	**51.13**

B.3 CANDIDATE-AVAILABILITY DIAGNOSTIC

Figure 2 evaluates uniform subsets drawn without replacement from frozen 128-candidate MATH-500 pools. For each of Llama, MetaMath, and Muggle, all five scorers are evaluated on the same 500 questions, with generator means weighted equally. All estimates, including SCM's, are computed from the final candidate-level outputs. The diagnostic therefore shares the BoN evaluation's question denominator, while averaging over subsets instead of ordered first-N prefixes.

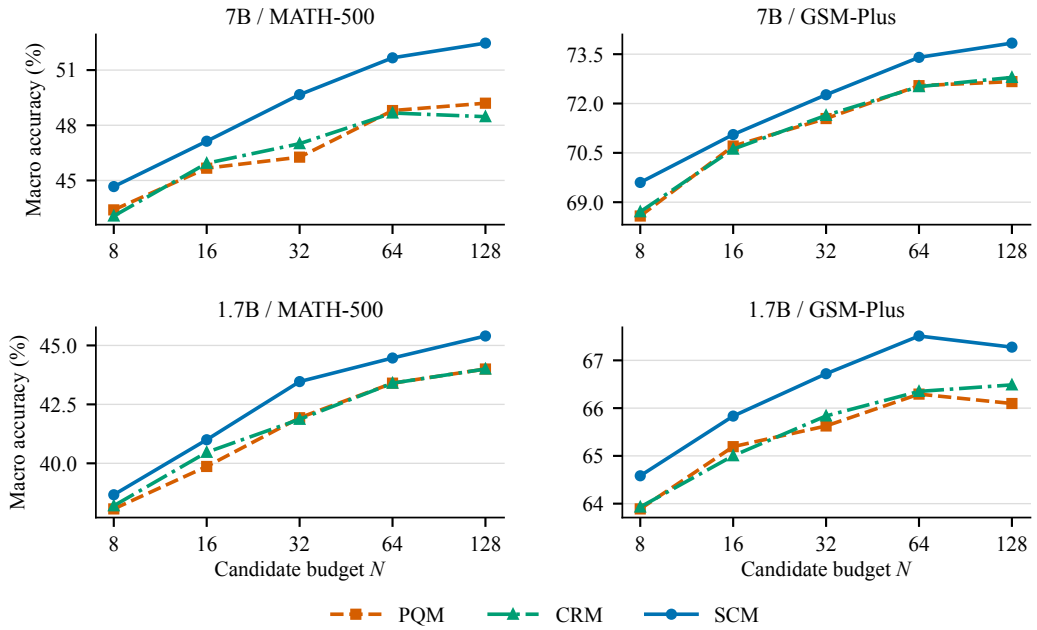


Figure 5: Generator-averaged BoN accuracy across the five evaluation budgets. Each point equally averages the three generator pools; lines connect the reported points only.

Table 8: Complete internal MATH-500 curves for Muggle-Math-13B. All values are accuracies (%) except the normalized AUBC summary. $N = 1$ is verifier independent and excluded from AUBC. Bold and underline indicate the two highest distinct values within each $N \geq 8$ column and AUBC.

Condition	@1	@8	@16	@32	@64	@128	AUBC
Step-only	18.4	34.8	37.8	40.0	41.6	42.2	39.48
Set-only	18.4	**37.0**	**40.2**	<u>41.2</u>	<u>43.4</u>	44.2	<u>41.35</u>
Uniform pairs	18.4	35.0	38.4	40.6	42.4	43.4	40.15
Full-pool pair	18.4	34.8	38.2	40.0	42.4	43.0	39.88
Hardest negative	18.4	35.0	38.4	40.6	42.2	43.2	40.08
Finite subsets	18.4	34.4	38.6	40.8	43.0	<u>44.4</u>	40.45
Matched top mass	18.4	35.2	38.6	40.8	42.8	43.6	40.40
SCM	18.4	<u>36.6</u>	<u>39.4</u>	**42.4**	**44.6**	**45.8**	**41.90**

For scorer m, let $c_{q,m}$ indicate a correct full-pool decision and let $r_{q,m}(N) = \mathbb{P}(\mathsf{M} \mid q, N)R^{\mathrm{wor}}_{q,m,N}$ be the joint probability that the subset contains a correct candidate and selection is incorrect. This contribution is zero when mixed subsets are impossible. With $\langle\cdot\rangle$ denoting the equal-generator average over all 500 questions per generator, the two plotted contributions are

$$H_m(N) = \langle c_{q,m} r_{q,m}(N)\rangle, \qquad F_m(N) = \langle (1 - c_{q,m}) r_{q,m}(N)\rangle. \tag{38}$$

Thus total expected error is $C(N) + H_m(N) + F_m(N)$, where $C(N)$ is the shared no-correct-candidate contribution from Equation 2. Both segments use the full 500-question denominator, while full-pool status is scorer-specific. Because $C(N)$ depends only on candidate labels, omitting it preserves accuracy differences between scorers at a fixed budget. Across budgets, however, $C(N)$ also changes, so bar-height changes alone do not measure changes in overall error.

SCM and the external PRMs use this same decomposition of their final candidate-level results. The figure's data file accompanies the source package.

B.4 Validation separation and test selection

Figure 6 compares complete separation on 660 mixed validation groups with test selection performance. Complete separation concerns all positive–negative rankings within a validation group,

whereas test AUBC measures answer selection across budgets. These trajectory-ranking diagnostics complement step-level error detection benchmarks such as ProcessBench and PRMBench (Zheng et al., 2025; Song et al., 2025).

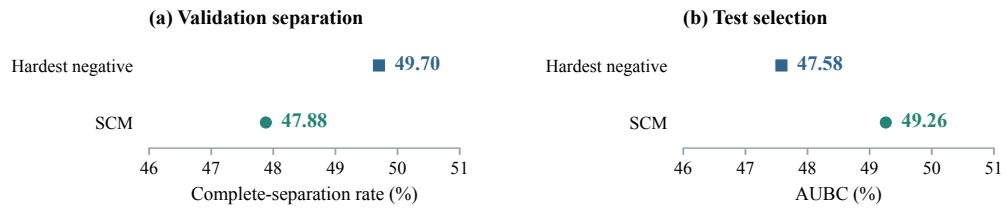


Figure 6: **Complete separation and selection performance.** (a) Fraction of the 660 mixed validation groups with every correct trajectory ranked above every incorrect one. (b) Test MATH-500 AUBC averaged over three generators. The two panels report validation and test summaries, respectively.

B.5 Robustness to Candidate-Availability Laws

SCM uses a uniform with-replacement law during training, whereas the main BoN evaluation uses ordered first-N prefixes and Figure 2 uses without-replacement subsets. To examine how this choice affects relative performance, we fix the three generators' MATH-500 pools, each containing 128 candidates for every one of the 500 questions, and the final 7B PQM, CRM, and SCM scores. Each method retains its main-experiment score aggregation and deterministic tie-breaking; no model is retrained. Overall accuracy and AUBC include all 500 questions per generator, with macro AUBC giving the generators equal weight. Conditional risks and champion-law discrepancies are evaluated only on mixed pools containing both correct and incorrect trajectories, where the conditioning event has positive probability.

We compare the WR and WOR joint champion distributions at each budget using d_{sel} and total variation from Appendix A.3. The d_{sel} calculation uses the final-row maxima of the dynamic program in Equation 27, after all positive-rank rows have been processed. Table 9 reports mean discrepancies on this mixed-pool support as probability values.

Table 9: WR/WOR joint champion-law discrepancy on mixed MATH-500 pools. Both columns are mean probability values, not percentages; d_{sel} uses the terminal-state dynamic-program result.

N	Mean d_{sel}	Mean TV
8	0.01	0.01
16	0.02	0.03
32	0.05	0.07
64	0.12	0.16
128	0.37	0.57

Table 10 compares ordered-prefix selection with expected selection performance under uniform WOR and WR laws over budgets $N \in \{8, 16, 32, 64, 128\}$. AUBC uses Equation 18; the ordered-prefix row reproduces the matched 7B results in Table 1.

Table 10: MATH-500 macro AUBC (%) under three candidate-availability laws. Each generator includes all 500 questions, and the three generators receive equal weight. Bold marks the highest value in each row.

Availability law	PQM	CRM	SCM
Ordered prefix	46.76	46.84	**49.26**
Uniform WOR	46.15	46.17	**48.71**
Uniform WR	45.78	45.82	**48.29**

SCM achieves the highest macro AUBC under all three protocols, improving over CRM by 2.42–2.54 pp and over PQM by 2.50–2.56 pp, with differences computed before rounding. Although the WR/WOR discrepancy increases with budget, SCM's relative advantage remains stable across these availability protocols on the fixed candidate pools.

The role of the candidate distribution also appears in reward learning from best-of-N preferences, where the sampling budget and base distribution affect the learned target (Pukdee et al., 2026). Our comparison holds the scorer fixed and changes the evaluation law, isolating a different consequence of candidate availability.

C Implementation and Search Details

C.1 Training data and supervision

The shared SCM recipe uses 200,000 trajectories in 12,428 question groups from RLHFlow/Mistral-PRM-Data and RLHFlow/Deepseek-PRM-Data. These released datasets contain reasoning trajectories for problems from the MATH and GSM8K training sets, with an aligned process label for each step.

Label construction. RLHFlow's annotation protocol follows Math-Shepherd-style hard estimation (Wang et al., 2024): from each intermediate reasoning prefix, it samples 16 continuations and assigns + if at least one reaches a correct final answer, and − otherwise. Thus the label records sampled continuation success from that prefix. SCM reuses these released annotations.

Step supervision in SCM. We map + to $b_{y,t} = 1$ and − to $b_{y,t} = 0$, training the scalar step head with binary cross-entropy. Let $\mathcal{I}_{\text{step}}$ contain the valid labeled step positions across the training batch. With $p_{y,t} = \text{sigmoid}(z_{y,t})$ from Equation 3,

$$\mathcal{L}_{\text{step}} = -\frac{1}{|\mathcal{I}_{\text{step}}|} \sum_{(y,t)\in\mathcal{I}_{\text{step}}} \left[b_{y,t} \log p_{y,t} + (1 - b_{y,t}) \log(1 - p_{y,t}) \right].$$

Each step retains its own process label, while released terminal labels define the correct and incorrect trajectory sets $\mathcal{P}_q$ and $\mathcal{N}_q$. Both losses update the same scorer: step BCE supervises individual positions, and the trajectory term in Equation 16 compares their mean scores with SCM weights. Single-outcome groups still supply step supervision. All author-trained models, including every ablation, are evaluated at the final checkpoint of their training runs.

Process-level signals also support policy training through stepwise preference optimization (Lai et al., 2024), implicit process rewards (Cui et al., 2026), and min-form credit assignment (Cheng et al., 2025). Here the process and terminal labels jointly train a verifier whose trajectory score is used for candidate selection.

C.2 Backbones and adaptation

Table 11: Author-trained model configurations. Methods sharing a backbone use the same training data, adaptation regime, and training budget. All models and ablations use their final checkpoints.

Backbone	Author-trained methods	Adaptation regime
Qwen2.5-Math-7B-Instruct	SCM, PQM, CRM, and ablations	Rank-64 LoRA
Qwen3-1.7B	SCM, PQM, CRM	Full-parameter training
Qwen3-0.6B	SCM	Full-parameter training

The 7B rank-64 adapter targets attention and feed-forward projections with alpha 128 and dropout 0.05, while the backbone remains frozen. The Qwen3 models use full-parameter training. Downloaded PRMs are evaluated as released.

SCM uses the same training configuration across all three backbones: 1,554 optimizer steps, eight question groups per step, bfloat16 (bf16) precision, a maximum trajectory length of 4,096 tokens, and AdamW (Loshchilov & Hutter, 2019). The trainable transformer parameters (LoRA adapters at 7B and the full backbone for Qwen3) and scalar head use learning rates of 10^{-4} and 3×10^{-4}, respectively, with a 78-step linear warmup followed by linear decay. The internal studies use the 7B recipe. Loss ablations remove either step BCE or the trajectory term; comparison-weight controls retain both losses and vary W.

Trajectory-loss schedule. In Equation 16, $\lambda(t)$ denotes the coefficient applied to the trajectory loss. For all three SCM backbones with $T = 1{,}554$ optimizer steps, it is

$$\lambda(t) = \lambda_{\max} \operatorname{clip}\left(\frac{t - 0.1T}{0.1T}, 0, 1\right), \qquad \lambda_{\max} = 1.0. \tag{39}$$

Thus the trajectory term is zero through step 155, becomes nonzero at step 156, and reaches 1.0 at step 311. The continuous ramp spans 10%–20% of training, or 155.4 steps.

Set-only training schedule. Set-only removes step BCE and uses an independent ramp:

$$\lambda_{\text{set-only}}(t) = \operatorname{clip}\left(\frac{t}{0.2T}, 0, 1\right).$$

The coefficient increases linearly from zero to one during the first 20% of training and remains one for the remaining 80%. Thus trajectory supervision is active throughout the initial ramp for $t > 0$.

C.3 Final checkpoints and public sources

Author-trained models and ablations use final checkpoints; downloaded PRMs use released checkpoints. Held-out groups provide the separation diagnostic (Appendix B.4). Training data come from RLHFlow/Mistral-PRM-Data and RLHFlow/Deepseek-PRM-Data. The backbones are Qwen2.5-Math-7B-Instruct, Qwen3-1.7B, and Qwen3-0.6B.

C.4 Best-of-N evaluation protocol

Llama3-70B-Instruct (Grattafiori et al., 2024), MetaMath-Mistral (Yu et al., 2024), and Muggle-Math-13B (Li et al., 2024a) each provide 128 frozen ordered candidates per question: 500 MATH-500 questions and 2,400 GSM-Plus questions. At $N \in \{8, 16, 32, 64, 128\}$, scorers select the highest-scoring first-N candidate using shared labels, order, and deterministic tie-breaking. SCM averages sigmoid scores at valid steps (Equation 3); comparison PRMs retain their aggregation rules. Accuracy includes all questions; macro AUBC equally averages the three generator AUBCs (Equation 18), with differences computed before rounding. Internal ablations reuse this MATH-500 protocol. The WOR diagnostic also retains all 500 questions per generator (Appendix B.3).

C.5 OlympiadBench evaluation protocol

BoN@8 uses the 674-problem English text-only mathematics subset of OlympiadBench (He et al., 2024), `olympiadbench_oe_to_maths_en_comp`. The seven verifiers share eight frozen Qwen3-4B responses per problem (5,392 total), ordered by `sample_index` from 0 to 7. The generator receives neither reference answers nor worked solutions.

Qwen3-4B uses its chat template with `enable_thinking=False` and a step-by-step reasoning prompt (Wei et al., 2022):

```
You will be given a problem.  Please reason step by step,
and put your final answer within \boxed{}:
[problem]
```

Sampling uses temperature 0.8, top-p 0.95, and `skip_special_tokens=True`, with eight candidates per request. The context limit is 4,096 tokens; completion length ranges from 128 to $\min(2048, 4096 - L_{\text{prompt}} - 8)$ tokens, where L_{prompt} is the prompt length.

Correct counts out of 674 are: Qwen2.5-Math-PRM-7B 243, SCM 241, ReasonFlux-PRM 234, PQM 233, CRM 229, SCAN-Pro 217, and PRM800K 214. Figure 4 rounds these ratios.

D Beam-Search Configuration

Search can compare partial solutions (Yao et al., 2023; Guan et al., 2025) or guide token decoding with rewards (Khanov et al., 2024). Table 3 uses the same verifier checkpoints as BoN. The root budget N_{root} counts first-layer candidates, distinct from beam width and total expansions. All verifiers

share the listed settings. The maximum action-sampling count $N_{\text{root}} + (T_{\max} - 1)Bf$ is 128 or 256 before early completion. MATH-500 and Gaokao2023 contain 500 and 383 questions, respectively.

Table 12: Beam-search configuration.

Parameter	$N_{\text{root}} = 8$	$N_{\text{root}} = 16$
Beam width	4	8
Children per surviving beam	2	2
Maximum reasoning steps	16	16
Sampling temperature	0.7	0.7
Maximum tokens per step	2,048	2,048
Random seed	42	42
Maximum action samples	128	256