DO PROCESS REWARD MODELS FAIL TO DETECT ERRORS, OR FAIL TO DECIDE? A CONTROLLED STUDY OF OPERATING POINTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Process reward models (PRMs) score intermediate reasoning steps, and judging the steps of a given solution turns those scores into decisions at a threshold. When a PRM misjudges steps, it is unclear whether its scores lack the information to detect errors or whether its decision boundary is simply misplaced. We separate these explanations with a controlled study that pairs a frozen-score audit of released PRMs on ProcessBench, PRMBench, and MedPRMBench with training interventions on 6,144 PRM800K chains that change either the clean-chain mixture or, on otherwise byte-identical inputs, only the class weights in the loss. We find that (1) released PRMs often discriminate far better than they decide: ReasonFlux-PRM and Qwen-Math reach nearly equal best shared-threshold ProcessBench Macro-F1 (73.36 vs. 73.50) yet differ by 28.35 points at the zero-margin boundary, which flags too few errors in 19 of 21 model–benchmark cells; (2) the training prior sets the logit origin rather than the ranking: loss-only interventions shift the operating point as the weighted-BCE optimum predicts (slope 0.987, R^2=0.953), shrink the best-to-fixed gap in all 12 paired Qwen3.5-4B runs, and reproduce 83–110% of the gap reduction from changing the mixture, while reverse weights reverse the effect; and (3) a source-set operating point transfers to a related benchmark but only partly to one with a higher error rate. Fixed-boundary evaluation thus conflates two failures that call for different fixes.

1 INTRODUCTION

Process reward models (PRMs) score the intermediate steps of a reasoning chain (Uesato et al., 2022; Lightman et al., 2024; Wang et al., 2024). Their scores rerank sampled solutions, guide decoding, and supply dense rewards for policy optimization (Lightman et al., 2024; Wang et al., 2024; Agrawal et al., 2026). Increasingly, PRMs are also asked to judge the steps of a *given* solution: to flag erroneous steps for scalable oversight (Zheng et al., 2025), or to locate the first error, which lets test-time search branch from the last valid prefix (Chitale et al., 2026). ProcessBench (Zheng et al., 2025), PRMBench (Song et al., 2025), and MedPRMBench (Wu et al., 2026) evaluate this role, the last in a safety-critical medical domain. Judging, unlike ranking, requires a decision: every flag, gate, or truncation compares a score with a boundary.

Measured this way, PRMs look like weak judges. On ProcessBench they fail to generalize beyond GSM8K- and MATH-level problems and trail prompted critic models (Zheng et al., 2025); PRMBench and MedPRMBench report significant weaknesses in error detection (Song et al., 2025; Wu et al., 2026). Yet each such number reads a PRM at one boundary, its argmax or emitted score 0.5 or a threshold tuned on a benchmark subset, and a low score there has two explanations. The scores may fail to separate erroneous from correct steps, a failure to *detect*; or they may separate them well around another boundary, a failure to *decide*. The two can come apart completely: adding a constant to every score logit preserves every ranking, yet can change every fixed-boundary decision (proposition 1). Indeed, two released PRMs with nearly identical step AUROC and best shared-threshold ProcessBench score differ by 28.35 points at the fixed boundary (fig. 1a). The distinction matters because the failures call for different fixes: better data, labels, or architectures for the first; a moved boundary, and knowledge of which training choice placed it, for the second. Reporting only the fixed boundary conflates them, in leaderboards and in the diagnoses built on them.

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.

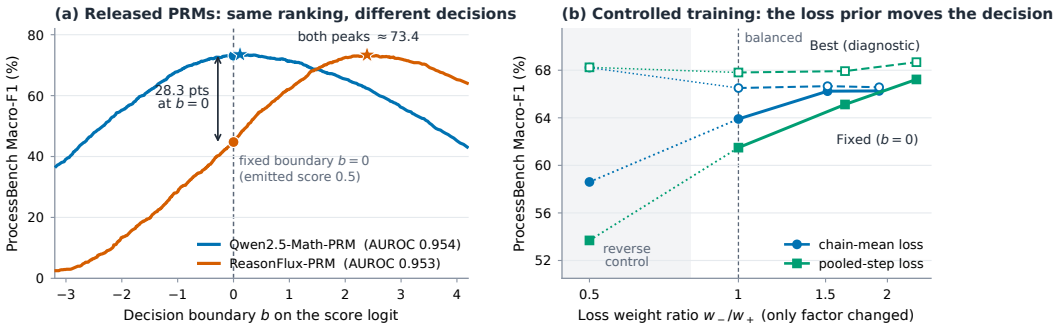


Figure 1: **Discriminating is not deciding.** (a) Frozen ProcessBench scores of two released PRMs at every shared boundary b: nearly equal step AUROC and best shared-threshold Macro-F1 (stars), yet 28.35 points apart at b=0 (circles); ReasonFlux peaks 2.4 logits away. (b) On our 4B testbed, changing only the loss ratio w_-/w_+ moves Fixed (filled) by up to 13.5 points and Best (open) by at most 1.7; ratio 0.5 (seed 42) reverses it. Forward points: mean of seeds 42/77/123. Best uses test labels (diagnostic only).

We therefore ask three research questions. **RQ1:** *When released PRMs perform poorly at their fixed boundary, do they lack step-level discrimination, or is that discrimination read at the wrong operating point?* **RQ2:** *What determines where a PRM's operating point lands, and can training move it without changing what the scores discriminate?* **RQ3:** *How far does an operating point chosen on one distribution travel to another?*

Existing work leaves these questions open because it treats weak step judgment as a problem of the scorer, and its remedies change what the scores are. Noisy Monte-Carlo labels are filtered or denoised (Zhang et al., 2025b; Ding et al., 2025; Xie et al., 2026); supervision targets become step rankings, implicit or trajectory-aware rewards, or error-typed labels (Li & Li, 2025; Yuan et al., 2024; Zou et al., 2025; Pala et al., 2025); readers generate critiques, score in both directions, or probe hidden states (Zhao et al., 2026; Khalifa et al., 2026; Zhang et al., 2026; Ni et al., 2026); and imbalanced supervision, linked to overcrediting plausible but incorrect steps, is countered by contrastive retraining (Agrawal et al., 2026). These remedies are typically evaluated at a fixed, subset-tuned, or unreported boundary, so a gain may come from better discrimination or from a moved boundary. Where boundaries are varied, the effect is treated as a nuisance or attributed to the scorer: ProcessBench reports per-subset thresholds for two scalar PRMs only as an upper bound (Zheng et al., 2025), FreePRM finds its F1 optimum at a threshold of 0.9 (Sun et al., 2025), and R-PRM reads a baseline's threshold shift as weaker generalization (She et al., 2025). Closer lines of work stop short of the distinction. Calibration makes scores reliable probabilities (Guo et al., 2017; Park et al., 2025; Ma et al., 2026), yet a calibrated score need not be F1-optimal at 0.5 (Lipton et al., 2014); a generative verifier's strictness can be steered through its hidden states, but what sets it in training is left open (Zhou et al., 2026). Classical results predict that the class prior shifts every logit by a constant (Saerens et al., 2002; Menon et al., 2021) and that a good representation can be held back by its classifier (Kang et al., 2020); to our knowledge they have not been used to separate the two failures of PRMs or tested under controlled PRM training (further related work: section J).

We answer these questions by pairing observation with intervention (section 2). Released PRMs differ in backbone, data, labels, loss, and score interface, so they can establish the phenomenon but not its cause. For RQ1, a frozen-score audit reads seven released PRMs on ProcessBench, PRMBench, and MedPRMBench at every shared boundary, recording fixed-boundary performance (Fixed), the best performance one shared threshold recovers (Best, a test-label diagnostic), step AUROC, and the flag rate relative to the error rate. For RQ2, Qwen3.5-4B PRMs share backbone, 6,144 training chains, initialization, example order, and seeds, and differ only in the loss's class weights, with a data-mixture arm for comparison (table 1); the weighted-BCE optimum (Menon et al., 2021) predicts the boundary shift, and reversed weights the opposite; we repeat the contrast at 9B and on Llama-3.1-8B. For RQ3, we carry thresholds chosen on source validation or on k target labels to PRMBench, MedPRMBench, and AgentProcessBench (Fan et al., 2026). Result 1 first checks that the controlled testbed is competitive (section 3). Our main findings are:

1. **Released PRMs often discriminate better than they decide (RQ1).** In 18 of 21 model–benchmark cells a different shared boundary gains at least 2 points; step AUROC orders Best (Spearman 0.96 on two benchmarks) far better than Fixed. In 19 cells the zero margin flags too few errors, an offset carried mostly by the PRM (77% of its variance); read at the target error rate, 17 cells come within a point of Best (section 4).
2. **The training prior sets the logit origin, not the ranking (RQ2).** Loss-only changes move the operating point as predicted (slope 0.987, R^2=0.953), shrink the best-to-fixed gap in 12/12 pairs, and move Best by only $+0.30$ points; reversed weights reverse the gap, and weighting alone reproduces 83–110% of a data mixture's gap reduction (section 5). The shift keeps its direction at 9B, but the Llama pilot shows no pure shift (section 7).
3. **A source-set operating point travels only partly (RQ3).** It nearly recovers the PRM-Bench optimum but leaves 4.46 points on MedPRMBench, largely because that target has a higher error rate (section 6).

A low fixed-boundary score is therefore not evidence of weak discrimination, and a remedy that raises Fixed but not Best has moved the boundary rather than improved discrimination. PRM evaluations should report both, marking any use of test labels.

2 A Controlled Framework for Separating Discrimination from Decision

2.1 Scores, readers, and the operating point

For a reasoning chain $x = (q, s_1, \ldots, s_T)$, a causal PRM emits $r_i = f_\theta(q, s_{\le i}) \in [0, 1]$. A shared threshold τ labels step i correct when $r_i \ge \tau$; PRMBench and MedPRMBench score these step decisions directly. ProcessBench targets $y \in \{0, \ldots, T\}$, where 0 is a clean chain and $k > 0$ its first error, and reads the first crossing

$$\hat{y}(\tau) = \min\{i : r_i < \tau\}, \qquad \min \emptyset := 0. \tag{1}$$

For binary-token verifiers with logits (a_i^+, a_i^-), the margin $z_i = a_i^+ - a_i^-$ makes the emitted argmax the rule $z_i \ge 0$, i.e., $\sigma(z_i) \ge 0.5$. Scalar and binary-token PRMs thus share one margin boundary $z_i \ge b$ with zero-margin reference $b = 0$. We call the boundary at which a PRM's scores are best separated for a task its *operating point*.

Proposition 1 (A score shift moves fixed decisions)**.** *For any constant c, replacing every margin by $z_i' = z_i + c$ preserves all pairwise orderings and step AUROC. It shifts the whole shared-boundary curve by c, while a fixed zero-margin reader stays at $b = 0$ and can decide differently.*

Proof. $z_i' > z_j'$ iff $z_i > z_j$, and a decision at b on z' is the decision at $b - c$ on z. □

Discrimination versus decision. Let $F(b)$ be the benchmark metric at boundary b. We report

$$F^\star = \max_{b \in \mathcal{B}} F(b), \qquad \mathcal{R} = F^\star - F(0), \tag{2}$$

where $\mathcal{B}$ contains every observed score breakpoint, its successor float, and 0.5. *Best* ($F^\star$) measures the discrimination available to one shared boundary; *Fixed* ($F(0)$) is the label-free zero-margin decision; their *gap* $\mathcal{R}$ is the performance discarded by the operating point. Best uses test labels and is diagnostic only. Observed-prefix step AUROC gives a threshold-free check of discrimination, and Radius@ϵ, the largest r with $F(b) \ge F(0) - \epsilon$ for $|b| \le r$, gives local sensitivity; a broad poor plateau is stable but not useful, so radius is read only with Best and the gap. For ProcessBench, an error at k is located exactly when $r_k < \tau \le \min_{j<k} r_j$; if an error scores above an earlier step, no shared threshold recovers the chain, which is a discrimination rather than an operating-point failure. Because optima of ratio metrics depend on the evaluation distribution (Lipton et al., 2014), we compare scores only within a benchmark.

2.2 Experimental framework

Benchmarks. ProcessBench (3,400 chains; four-subset first-error Macro-F1), PRMBench (6,211 records; official PRMScore), and MedPRMBench (6,500 corrupted medical chains; step Macro-F1; Wu et al., 2026). Cohorts and metrics are detailed in section A.

Table 1: **Only loss weights change within the main pairs.** The separate data-mixture arm changes examples to test whether weighting reproduces the mixture's gap reduction.

Held fixed within a pair	Qwen3.5-4B binary MLP; the same 6,144 PRM800K chains (50% clean), example order, initialization, optimizer, training length, seed, and loss reduction.
Changed in the pair	Loss ratio w_-/w_+: baseline 1 to 1.514 or 1.925 (chain mean), 1.642 or 2.289 (pooled steps); 0.5 reverses direction. Each reduction has its own baseline.
Data-mixture check	Clean-chain share $50\% \rightarrow 30\%$ or 15% at 6,144 chains; selected examples change. Compare its gap reduction with a prior-matched loss-only run.

Frozen-score audit. We replay seven released scalar PRMs with their released prompts and score adapters, keep full-precision scores, and verify reference evaluators at six thresholds before admitting a curve. The 0.5 convention is induced by emitted scores and is not necessarily author-endorsed (table 10); the audit alleges no misuse of ranking-only PRMs. Delta-reward, generative-verdict, and additional scalar interfaces are appendix checks (section B).

Controlled training testbed. The main controlled comparison fine-tunes a Qwen3.5-4B binary-MLP PRM on PRM800K step labels. Table 1 separates two interventions: loss-only pairs change w_-/w_+ on byte-identical examples, whereas the composition arm changes the selected chains. Forward ratios match the effective prior of 30%- or 15%-clean compositions; ratio one is each reduction's paired baseline and ratio 0.5 is a pre-specified reverse-direction control. Uncertainty uses 1,000 hierarchical paired bootstrap replicates over seeds and stratified ProcessBench chains. For transfer, one threshold is selected on saved source-validation predictions before any target replay.

3 Result 1: Is the Controlled Testbed a Credible PRM?

A controlled study is only informative if its models are not weak outliers: an operating-point effect in a poor PRM could simply reflect poor scores. We therefore check the testbed before using it.

Setting. We train the same Qwen3.5 binary-MLP PRM used in the controlled interventions on 20K PRM800K chains with pooled-step BCE at 4B and 9B, read it at the zero-margin boundary, and place it beside local replays of released PRMs under their emitted conventions (table 2).

Observations. With no threshold tuning, the 9B model reaches 75.2 ProcessBench Macro-F1 and 68.5 PRMBench PRMScore, above the strongest released replay (Qwen2.5-Math-PRM-7B: 73.1 and 65.5); it leads on three of four ProcessBench subsets and on every PRMBench group average. The gain comes from locating errors (ProcessBench Err. 61.5–76.8 versus 54.7–71.0), while it accepts fewer clean chains. The 4B model is within 4 points on ProcessBench and above every released replay on PRMBench. On MedPRMBench the 9B model reaches 70.9 step Macro-F1 against at most 70.4 for released PRMs, and the 4B model 69.0 (tables 7 and 8).

Takeaway. *The controlled testbed is a competitive PRM across three domains, so effects measured on it are not artifacts of weak scores.*

Discussion. These are unmatched comparisons: backbones, training data, and interfaces differ, and published generative verifiers use far larger inference budgets (e.g., GenPRM-7B reports 75.2 Pass@1 and 80.5 Maj@8 on ProcessBench; Zhao et al., 2026). The interventions train this PRM on 6,144 chains; fewer chains lower Best but do not move the operating point (section G).

4 Result 2: Do Released PRMs Fail to Discriminate, or Fail to Decide?

Setting. For seven released scalar PRMs we freeze the scores on each benchmark and read them at every shared boundary, recording Fixed, Best, their gap, and observed-prefix step AUROC (fig. 2;

Table 2: **Prerequisite check: the controlled testbed is a competitive PRM.** Public rows are complete local replays of released checkpoints at their emitted-score 0.5 convention (Skywork: its −0.05 reward-drop rule); ours are means over seeds 42/77/123 (sample SD on the headline column). Rows differ in backbone, data, and interface, so comparisons are unmatched; no row tunes a threshold on test labels. Bold: column best; underline: best released PRM (F1 and average columns).

(a) ProcessBench. Err./Cor.: exact error localization / correct-chain acceptance; F1: their harmonic mean; Avg.: subset-macro F1.

	GSM8K			MATH			OlympiadBench			OmniMath			
PRM	Err.	Cor.	F1	Err.	Cor.	F1	Err.	Cor.	F1	Err.	Cor.	F1	Avg.
Public checkpoints (local replay)													
Qwen2.5-Math-PRM-7B	71.0	96.4	81.8	67.2	91.1	77.3	55.2	85.5	67.1	54.7	83.8	**66.2**	73.1
Qwen2.5-Math-7B-PRM800K	53.1	95.9	68.4	48.0	90.1	62.6	35.9	87.3	50.8	29.2	86.7	43.7	56.4
Math-Shepherd-7B	32.4	93.3	48.1	16.5	82.5	27.5	11.0	74.9	19.3	11.2	77.2	19.6	28.6
RLHFlow-Mistral-8B	34.3	98.4	50.9	21.9	72.4	33.6	8.3	44.0	14.0	9.7	44.0	16.0	28.6
RLHFlow-DeepSeek-8B	23.7	98.4	38.2	22.2	80.0	34.8	10.0	51.6	16.7	10.4	51.9	17.3	26.8
Skywork-PRM-1.5B	46.4	51.3	48.7	31.3	45.3	37.0	18.8	27.4	22.3	13.6	37.8	20.0	32.0
Skywork-PRM-7B	50.2	55.4	52.7	30.8	41.6	35.4	18.8	23.3	20.8	13.6	40.7	20.3	32.3
Universal-PRM-7B	0.0	100.0	0.0	0.0	100.0	0.0	0.0	100.0	0.0	0.0	100.0	0.0	0.0
Controlled testbed (ours): PRM800K, unweighted pooled-step BCE													
Ours Qwen3.5-4B pooled ($n = 3$)	66.8	95.7	78.7	68.2	83.5	75.1	59.7	71.1	64.9	53.0	66.0	58.7	$69.3_{\pm 0.5}$
Ours Qwen3.5-9B pooled ($n = 3$)	76.8	94.5	**84.7**	75.1	85.5	**80.0**	67.7	74.9	**71.1**	61.5	68.9	65.0	$\mathbf{75.2}_{\pm 0.5}$

(b) PRMBench. Published category protocol (MS.: positive F1; others: PRMScore); group Avg.: mean of displayed categories; Overall: PRMScore over all non-correct records. Abbreviations in section B.

	Simplicity			Soundness					Sensitivity				
PRM	NR.	NCL.	Avg.	ES.	SC.	DC.	CI.	Avg.	PS.	DR.	MS.	Avg.	Overall
Public checkpoints (local replay)													
Qwen2.5-Math-PRM-7B	49.0	55.0	52.0	71.8	67.4	66.4	78.4	71.0	57.7	69.0	99.8	75.5	65.5
Qwen2.5-Math-7B-PRM800K	48.5	47.7	48.1	62.1	59.6	58.7	68.4	62.2	52.8	63.9	99.8	72.2	58.3
Math-Shepherd-7B	44.6	50.8	47.7	50.0	45.3	41.9	48.4	46.4	47.6	49.4	86.0	61.0	47.6
RLHFlow-Mistral-8B	46.3	47.2	46.8	56.4	55.3	54.2	63.7	57.4	51.5	56.3	98.1	68.6	54.4
RLHFlow-DeepSeek-8B	46.4	48.5	47.5	54.8	54.9	53.0	65.8	57.1	48.5	55.5	99.8	67.9	53.9
Skywork-PRM-1.5B	54.1	61.4	57.7	65.1	62.6	62.5	65.2	63.9	58.2	62.4	89.8	70.1	61.5
Skywork-PRM-7B	53.4	60.1	56.7	64.9	62.4	62.4	65.7	63.8	58.2	62.3	93.0	71.2	61.4
Universal-PRM-7B	46.5	42.1	44.3	44.5	47.1	46.4	46.8	46.2	44.6	45.4	100.0	63.3	45.7
Controlled testbed (ours): PRM800K, unweighted pooled-step BCE													
Ours Qwen3.5-4B pooled ($n = 3$)	52.0	64.5	58.2	71.0	69.9	69.9	74.7	71.4	60.9	68.9	99.6	76.5	$67.4_{\pm 0.1}$
Ours Qwen3.5-9B pooled ($n = 3$)	51.9	64.8	**58.4**	72.7	71.3	70.5	76.6	**72.8**	62.0	70.4	99.5	**77.3**	$\mathbf{68.5}_{\pm 0.1}$

numbers in table 9). Varying only the boundary within a frozen checkpoint isolates the decision rule from everything the model has learned.

Observations. *The fixed boundary is rarely the best one.* In 18 of 21 model–benchmark cells a different shared boundary gains at least 2 points; the median gain is 9.6, 5.7, and 22.8 points on ProcessBench, PRMBench, and MedPRMBench (fig. 2a). *Discrimination, not the fixed score, tracks ranking quality.* Step AUROC orders Best almost perfectly across PRMs (Spearman 0.96 on ProcessBench and PRMBench) but orders Fixed less well (0.64, 0.75). Qwen-Math and ReasonFlux differ by 0.001 in AUROC and 0.14 points in Best, yet by 28.35 in Fixed, and Universal-PRM recovers 70.97 from 0.00. Fixed and Best leaderboards correlate at only 0.46, 0.71, and 0.18, and on MedPRMBench the top model changes. *The displacement is systematic and mostly a model property.* In $19/21$ cells $b^\star > 0$ (median $|b^\star| = 1.25$ logits): the zero margin flags too few errors. With $\rho = r/\pi$ the fraction of units rejected at $b{=}0$ over the target error prevalence (chains on ProcessBench, steps otherwise), exactly these 19 cells have $\rho < 1$; ReasonFlux flags MedPRMBench steps at 1% of their error rate and Universal-PRM flags nothing (table 3). 77% of the variance of $b^\star$ is attributable to the PRM, 1% to the benchmark, and 22% to their interaction (fig. 2b). *In rank terms the optimum sits near the error rate.* The prevalence quantile Q_π, which rejects exactly a π fraction of units, is within 1 point of Best in $17/21$ cells (mean shortfall from Best 0.88 points, against 15.73 for Fixed), although the best thresholds lie up to 5.5 logits apart. At Q_π, class counts match, a

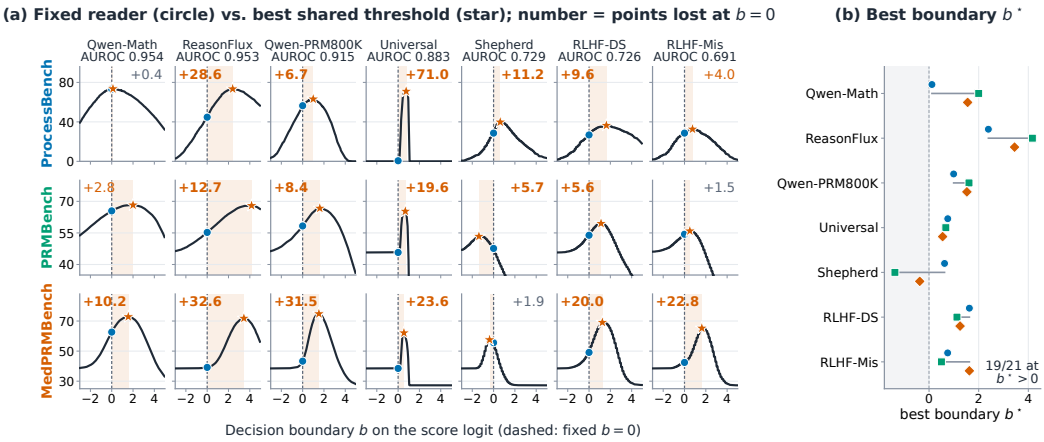


Figure 2: **The fixed boundary is almost never the best one.** (a) Each panel reads one released PRM's frozen scores on one benchmark at a shared logit boundary b. Circle: zero-margin reader (b=0, emitted score 0.5, a convention rather than always an endorsed rule; table 10); star: exact best shared threshold (test labels, diagnostic only); number: Best−Fixed (orange if ≥ 2); shading: displacement. Columns follow ProcessBench step AUROC. (b) $b^\star$ of all 21 cells, colored by benchmark: $19/21$ lie at $b^\star > 0$, clustered by PRM rather than benchmark. Larger panels: fig. 4.

precision–recall break-even for pointwise step Macro-F1; its near-optimality under PRMScore and ProcessBench's first-error metric is empirical (section G). The four misses (three on ProcessBench, one on MedPRMBench) all involve Math-Shepherd or the RLHFlow models, whose lower ProcessBench AUROC (0.69–0.73) and Best (32.6–39.8) mark a genuine discrimination limit rather than an operating-point failure; in 59–71% of the RLHFlow models' ProcessBench error chains, no shared threshold can locate the error.

Table 3: **In rank terms the best boundary sits at the error rate.** ρ: zero-margin rejection rate over the error prevalence π (chains on ProcessBench, steps otherwise; $\rho < 1$: under-flagging). Q_π: scores read at the quantile that rejects a π fraction; π and Best use test labels, so Q_π is diagnostic. Misspecifying π by $\pm 30\%$ raises the mean shortfall from 0.88 to about 5.5 points (section G).

	ProcessBench (π=0.65)				PRMBench (π=0.16)				MedPRMBench (π=0.37)			
PRM	ρ	Fixed	Q_π	Best	ρ	Fixed	Q_π	Best	ρ	Fixed	Q_π	Best
Qwen-Math	0.92	73.1	72.6	73.5	0.55	65.5	68.1	68.2	0.41	62.6	72.6	72.8
ReasonFlux	0.42	44.8	72.7	73.4	0.15	55.2	67.9	68.0	0.01	39.2	71.7	71.8
Qwen-PRM800K	0.74	56.4	62.9	63.1	0.32	58.3	66.4	66.7	0.05	43.3	74.3	74.8
Universal	0.00	0.0	70.9	71.0	0.00	45.7	65.1	65.3	0.00	38.5	62.1	62.1
Shepherd	0.67	28.6	38.5	39.8	3.09	47.6	53.1	53.4	1.62	55.6	57.2	57.5
RLHF-DS	0.90	26.8	30.9	36.4	0.37	53.9	59.0	59.5	0.20	49.1	68.6	69.1
RLHF-Mis	0.98	28.6	28.6	32.6	0.66	54.4	55.1	55.9	0.14	42.5	63.2	65.2

Takeaway. *A low fixed-boundary score does not imply weak discrimination. Released PRMs rarely decide at their best boundary, and they miss it in a systematic direction: the zero margin flags fewer errors than the target contains. The offset is carried mostly by the model; in rank terms, the optimum sits close to the target error rate. We next ask which training choice sets the offset.*

Discussion. Best, π, and Q_π use test labels and are diagnostics, not deployment results. Still, a threshold chosen on half of the chains loses only 0.22 points on the other half (at most 0.64), and AUROC corroborates Best without any threshold. Read at Q_π, chance-corrected agreement between PRMs on which units are wrong rises (mean Cohen's κ 0.25, 0.17, 0.09 $\rightarrow$ 0.53, 0.37, 0.42), and part of the positive-reward bias reported by PRMBench (Song et al., 2025) is an operating-point effect (section G). The variance shares are descriptive over seven PRMs and three benchmarks; the

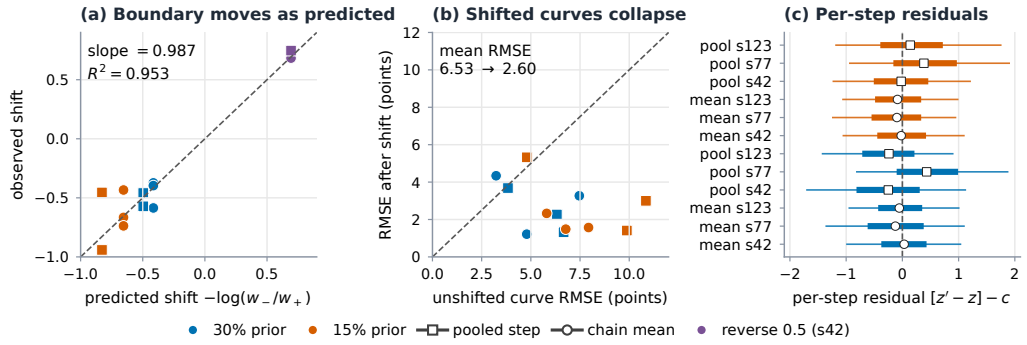


Figure 3: **Fixed-support diagnostics support a dominant but non-pure additive shift.** (a) The weighted-BCE prediction tracks the source-threshold displacement of forward and reverse interventions. (b) Applying the predicted shift lowers paired decision-curve RMSE in aggregate. (c) Per-step residuals remain after the shift (thin: 5–95%; thick: 25–75%; marker: median).

small benchmark share hides offsetting shifts in error rate and in scores between PRMBench and MedPRMBench. Because released models differ in every training choice, this audit establishes the phenomenon but not its cause.

5 Result 3: What Sets the Operating Point?

Result 2 shows that operating points drift but cannot say why. We now vary a single training factor on otherwise byte-identical inputs.

Setting and prediction. For step-supervised data, the loss is

$$\mathcal{L} = \sum_c \sum_{i \in \mathcal{O}_c} a_{ci} w_{y_{ci}} \ell(q_{ci}, y_{ci}), \qquad a_{ci} = \begin{cases} |\mathcal{O}_c|^{-1} & \text{chain mean,} \\ 1 & \text{pooled steps.} \end{cases} \tag{3}$$

Composition decides which labels occur, reduction their per-chain contribution, and w_y their explicit weight. For weighted BCE the pointwise population optimum is

$$\operatorname{logit} q^*(x) = \operatorname{logit} \Pr(y{=}1 \mid x) - \log(w_-/w_+). \tag{4}$$

so, with examples and reduction fixed, changing w_-/w_+ predicts an additive logit displacement (Menon et al., 2021): a shift of the operating point, not of the ranking. A finite model need not realize a pure constant shift, so we test the direction, its magnitude, ranking preservation, and a reverse intervention (section 2.2).

Observations. *The boundary moves as predicted.* All 12 4B forward interventions move the source-validation optimum in the predicted direction, and the predicted displacement $-\log(w_-/w_+)$ matches the observed one with slope 0.987 and $R^2 = 0.953$ across forward and reverse runs (fig. 3a). All 12 source optima move toward 0.5 (mean absolute distance $0.230 \to 0.112$). *Decisions change; discrimination barely does.* The best-to-fixed gap shrinks in 12/12 pairs by 3.21 points on average (hierarchical bootstrap -2.97, 95% CI $[-3.92, -1.88]$), and Fixed rises by 3.52 points (11/12), while Best changes by only $+0.30$ (table 4). Paired step ranks stay highly correlated (mean 0.948), and translating the baseline curve by the predicted shift cuts curve RMSE from 6.53 to 2.60 points (fig. 3b). *Reversing the prior reverses the effect.* The ratio-0.5 control increases the gap by 6.54 (chain mean) and 7.70 (pooled step) points with rank correlations of 0.962 and 0.952; the predicted translation again aligns the curves (RMSE $10.33 \to 2.28$ and $7.54 \to 2.39$). *Data composition acts largely through the same prior.* Changing the clean-chain mixture alters both support and label mass. Matching only its effective prior by loss weighting on unchanged 50%-clean support reproduces 83–110% of the gap reduction in group means, with the same sign in every seed. *Data volume does not move the boundary.* With the 15%-clean mixture and the loss fixed, growing

Table 4: **Fixed-support weighting changes Fixed, not Best.** Δ: weighted intervention minus its matched ratio-one baseline on ProcessBench, in points except event AUC (first-error position vs. all other positions, including the unlabelled post-error suffix; not step AUROC). Fixed: zero-margin Macro-F1; Best: test-label maximum; $\Delta\text{Gap} = \Delta\text{Best} - \Delta\text{Fixed}$, negative when weighting closes the gap. 4B forward groups: three paired seeds; reverse: seed 42; 9B: chain-mean 15%-prior forward and reverse only.

Scale	Intervention	ΔFixed	ΔBest	ΔGap	ΔEvent AUC
4B	30% prior, chain mean	+2.34	+0.16	−2.18	−0.0004
4B	30% prior, pooled step	+3.63	+0.12	−3.51	−0.0052
4B	15% prior, chain mean	+2.36	+0.07	−2.29	−0.0024
4B	15% prior, pooled step	+5.74	+0.87	−4.87	+0.0024
4B	All forward (12 pairs)	+3.52	+0.30	−3.21	−0.0014
4B	Reverse, chain mean (s42)	−6.79	−0.25	+6.54	−0.0022
4B	Reverse, pooled step (s42)	−6.99	+0.71	+7.70	−0.0048
9B	15% prior (3 pairs)	+0.85	−0.46	−1.31	−0.0055
9B	Reverse (s42)	−4.76	−0.56	+4.20	+0.0005

the training set from 1K to 20K chains (seed 42) raises Fixed from 59.4 to 68.8 and Best from 60.2 to 69.3, yet the source-validation threshold stays within 0.510–0.545; at 6,144 chains, raising the clean fraction to 30% and 50% moves it from 0.560–0.630 to 0.655–0.740 and 0.770–0.845 (three seeds). Without class weights, pooled-step runs flag close to the ProcessBench error rate (ρ between 0.94 and 1.09 at 4B and 9B); class-balanced weights ($w_-/w_+ \approx 6.8$) raise ρ to 1.25–1.53 (section G).

Takeaway. *The effective training prior sets the logit origin of a PRM, and thereby its operating point, while leaving what the scores discriminate largely unchanged. Most of the gap change from changing the mixture is reproduced by this same operating-point shift.*

Discussion. The shift is dominant but not pure: the mean within-pair per-step shift SD is 0.752 logits (fig. 3c), and weighting does not reproduce every radius or Best change of the most extreme mixture. The regression's R^2 is a within-design fit over a few weight ratios, not independent evidence across families. A smaller gap is also not a guarantee of a higher fixed score (one of 12 pairs). The mechanism is classical (Menon et al., 2021); what is new is that it quantitatively governs PRM first-error decisions with ranking preserved, unlike a score-quality account of imbalanced supervision (Agrawal et al., 2026; seed-level pairs: table 12).

6 Result 4: How Far Does a Source-Set Operating Point Travel?

Setting. For the 12 frozen 4B interventions we select one threshold on source validation before any target replay and compare it with the fixed reader on PRMBench and MedPRMBench (table 5), or select it from k labelled target records (table 15).

Table 5: **Source-validation thresholds transfer to PRMBench but leave a medical residual.** Values aggregate 12 frozen 4B loss-only interventions; the source threshold is fixed before target scoring.

	PRMBench	MedPRMBench
Δ(source−Fixed)	$+0.21$	$+3.39$
Better cells	9 / 12	11 / 12
Gap to Best	0.13	4.46

Observations. The source threshold is 0.13 below Best on PRMBench and 4.46 below it on MedPRMBench, though it beats the zero margin there by 3.39 in 11/12 cells. The step error rate rises from 0.16 to 0.37; with class-conditional scores fixed, this explains 1.17 of the 1.25-logit move of the best boundary. A diagnostic correction using target-label prevalence leaves 0.35 points (section G). Direct threshold tuning on 16–128 labelled records overfits on PRMBench (400 add just 0.16 points), while 16 records raise MedPRMBench by 6.82 over zero margin and 3.44 over source. Using those same 16 records only to estimate the target error rate scores 72.52 on MedPRMBench, versus 72.01 for direct tuning and 68.57 for source; PRMBench scores are 66.43, 65.38, and 66.59, respectively (12 runs, 100

record-disjoint calibration/evaluation splits; table 15). The labels mostly supply an error rate rather than a precise logit cutoff. On AgentProcessBench, a shared threshold removes about three quarters of the gap between two reductions across three seeds, with a seed-dependent remainder (table 16).

Takeaway. *Training moves a model-level offset that transfers to a related domain, while target prevalence and metric leave residuals. No fixed boundary is optimal everywhere.*

Discussion. The operating point is a property of a model, a benchmark, and a metric together (Lipton et al., 2014): a boundary that moves between targets is not by itself evidence of weaker generalization (She et al., 2025), and our results do not show that a balanced loss generalizes better, or that it beats a source threshold on an unweighted baseline. A label-free target-side probe is reported only as a hypothesis (section H), and a fixed-pool MATH-500 check changes the gate's sensitivity without a significant accuracy gain (section E).

7 Result 5: How General Is the Effect?

Setting. We repeat the 15%-prior chain-mean contrast with Qwen3.5-9B over three seeds plus a seed-42 reverse control, run a pre-specified single-seed three-arm pilot on Llama-3.1-8B, and apply the same decision profile to non-scalar interfaces.

Observations. *Scale.* At 9B the gap shrinks in all three seeds by 0.54–2.50 points (bootstrap -1.48, 95% CI $[-2.92, -0.33]$); the observed mean source shift is -0.766 against a prediction of -0.655, and the reverse control raises the gap by 4.20 points. Fixed Macro-F1 rises in only two seeds: in seed 77 the gap falls by 0.89 while Fixed and Best fall by 1.03 and 1.92 (table 13). *Backbone.* On Llama-3.1-8B the forward and reverse arms move step logits by median -0.797 and $+0.672$ against predictions of -0.655 and $+0.693$, but the forward arm also lowers prefix AUROC and loses 8.73 Best points, so its smaller gap is not a better reader (section C.1). *Interface.* The profile applies to a generated critique followed by a binary verdict: R-PRM-7B is nearly aligned on ProcessBench (63.86 Fixed, 64.01 Best), and delta-reward interfaces are profiled in section B.

Takeaway. *The direction of the prior-induced shift replicates across scale and backbone, but it behaves as a clean operating-point shift only when the training recipe preserves discrimination; the separation of Fixed and Best applies beyond scalar heads.*

Discussion. The single-seed Llama pilot cannot establish a family-level negative, but rules out a backbone-independent pure shift. Generative verifiers are not ranked against scalar PRMs.

8 Conclusion

We asked whether PRMs that misjudge steps fail to detect errors or fail to decide. Many released PRMs discriminate well but flag too few errors at their fixed boundary; others face a genuine discrimination limit. In matched runs, loss weights move the operating point as predicted while preserving ranking and reproduce most of the gap change from changing the mixture, and the offset travels to a related benchmark but only partly to one with a higher error rate. Training sets where a PRM decides; the target's error rate largely sets where it should. PRM evaluations should therefore report fixed-boundary performance beside the best one-threshold score, marking any use of test labels. Two quantities make the distinction cheap to report: the flag ratio ρ, which falls below one when the boundary under-flags, and the prevalence-quantile reader Q_π, which needs only the target error rate and comes within one point of Best in 17 of 21 audited cells, missing only where discrimination itself is limited. Whether the training prior likewise sets the strictness of generative verifiers (Zhou et al., 2026) is a natural next question.

Limitations. Best is a test-label diagnostic, not a deployable reader. The controlled evidence covers one backbone family at two scales, with a single-seed Llama pilot that does not show a pure shift. We have not evaluated a previously untouched target domain, and a fixed-pool MATH-500 check shows no significant downstream accuracy gain. Both ρ and Q_π take the target error rate from test labels; estimating it without labels remains a hypothesis (section H).

Ethics Statement

This study evaluates released PRMs and our own step classifiers on existing benchmarks and introduces no new human-subject data. MedPRMBench (Wu et al., 2026) is built from existing medical question-answering datasets with model-generated reasoning chains. PRM step judgments, including medical ones, are not validated for clinical decision-making; our results concern how such judgments are evaluated, not their deployment. Released models and datasets are used under their licenses.

Reproducibility Statement

Section 2.2 specifies benchmark cohorts, score interfaces, and the fixed-boundary reader. The appendix records dataset exclusions, released-model adapters, paired intervention settings, uncertainty estimates, and the scripts and frozen artifacts used to rebuild tables and figures (sections A to C and I); the anonymized supplementary material contains them, and its README maps each table and figure to its builder script and frozen output. All reported fixed scores use frozen predictions; target labels select only the explicitly diagnostic best shared thresholds, and source-transfer thresholds are chosen before target labels are read. The fixed-support study pins examples and seeds within each pair; its checkpoints and raw caches were not retained, so its tables are rebuilt from frozen per-run summaries, and section G reports a re-run on a rebuilt backbone. Exploratory target adaptation and the fixed-pool utility check are identified separately in the appendix.

AI Use Statement

Generative AI tools assisted with literature search and organization, experiment design, figure design, engineering and analysis code, drafting and editing, and checks of claims against recorded results. The authors are responsible for verifying the claims, numerical results, citations, code, and final submission; AI assistance does not replace that responsibility.

References

Aakriti Agrawal, Souradip Chakraborty, Armin Saghafian, Nihal Sharma, Rizal Fathony, Nam H Nguyen, C. Bayan Bruss, Amrit Singh Bedi, and Furong Huang. The hidden bias of process reward models: PRISM for rewarding the right reasoning. *arXiv preprint arXiv:2606.09078*, 2026.

Tyler Alvarez and Ali Baheri. Where does reasoning break? Step-level hallucination detection via hidden-state transport geometry. *arXiv preprint arXiv:2605.13772*, 2026.

Xinzhu Chen, Wei He, Huichuan Fan, Wenzhe Niu, Zhongxiang Sun, Xuanru Wang, Jiuchong Gao, Jinghua Hao, Renqing He, and Weijie Yu. Hidden states know where reasoning diverges: Credit assignment via span-level Wasserstein distance. *arXiv preprint arXiv:2604.23318*, 2026.

Rajiv Shailesh Chitale, Rahul Madhavan, Taneesh Gupta, Deepanway Ghosal, and Aravindan Raghuveer. Test-time scaling via error localization. *arXiv preprint arXiv:2607.21453*, 2026.

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.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

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*, 2025. arXiv:2509.16548.

Shengda Fan, Xuyan Ye, Yupeng Huo, Zhi-Yuan Chen, Yiju Guo, Shenzhi Yang, Wenkai Yang, Shuqi Ye, Jingwen Chen, Haotian Chen, Xin Cong, and Yankai Lin. AgentProcessBench: Diagnosing step-level process quality in tool-using agents. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2026. arXiv:2603.14465.

Artyom Gadetsky, Maxim Kodryan, Siba Smarak Panigrahi, Hang Guo, and Maria Brbić. Unsupervised process reward models. *arXiv preprint arXiv:2605.10158*, 2026.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017.

Jizhou Guo, Zhaomin Wu, Hanchen Yang, and Philip S. Yu. Mining intrinsic rewards from LLM hidden states for efficient best-of-N sampling. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2026. arXiv:2505.12225.

Tianyu Jia, Yue Fang, Hongxin Ding, Rihong Qiu, Zhibang Yang, Zhijing Wu, Xu Chu, Junfeng Zhao, and Yasha Wang. The weakest link tells it all: Outcome-supervised process reward modeling via learnable credit assignment. *arXiv preprint arXiv:2606.27739*, 2026.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*, 2020.

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.

Sanwoo Lee, Clive Bai, Hsiu-Yuan Huang, Kun Liang, Weijie Liu, and Yunfang Wu. Learning latent reasoning traces for scalar reward models end-to-end. *arXiv preprint arXiv:2607.29185*, 2026.

Jinyuan Li, Langlin Huang, Chengsong Huang, Shaoyang Xu, Donghong Cai, Yuyi Yang, Wenxuan Zhang, and Jiaxin Huang. Process rewards with learned reliability. *arXiv preprint arXiv:2605.15529*, 2026.

Wendi Li and Yixuan Li. Process reward model with Q-value rankings. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.11287.

Zhenwen Liang, Ruosen Li, Yujun Zhou, Linfeng Song, Dian Yu, Xinya Du, Haitao Mi, and Dong Yu. CLUE: Non-parametric verification from experience via hidden-state clustering. *arXiv preprint arXiv:2510.01591*, 2025.

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

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

Zachary C. Lipton, Charles Elkan, and Balakrishnan Narayanaswamy. Optimal thresholding of classifiers to maximize F1 measure. In *Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, pp. 225–239, 2014. doi: 10.1007/978-3-662-44851-9_15.

Zachary C. Lipton, Yu-Xiang Wang, and Alexander J. Smola. Detecting and correcting for label shift with black box predictors. In *International Conference on Machine Learning (ICML)*, 2018.

Rachel Ma, Dylan Hadfield-Menell, and Kristjan Greenewald. Distributional process reward models: Calibrated prediction of future rewards via conditional optimal transport. *arXiv preprint arXiv:2605.06785*, 2026.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations (ICLR)*, 2021.

Jingwei Ni, Ekaterina Fadeeva, Tianyi Wu, Mubashara Akhtar, Jiaheng Zhang, Elliott Ash, Markus Leippold, Timothy Baldwin, See-Kiong Ng, Artem Shelmanov, and Mrinmaya Sachan. Efficient test-time scaling of multi-step reasoning by probing internal states of large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11667–11689, 2026. doi: 10.18653/v1/2026.acl-long.536.

Tej Deep Pala, Panshul Sharma, Amir Zadeh, Chuan Li, and Soujanya Poria. Error typing for smarter rewards: Improving process reward models with error-aware hierarchical supervision. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 14940–14954, 2025. doi: 10.18653/v1/2025.findings-emnlp.807.

Young-Jin Park, Kristjan Greenewald, Kaveh Alimohammadi, Hao Wang, and Navid Azizan. Know what you don't know: Uncertainty calibration of process reward models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 43479–43522, 2025. doi: 10.52202/085713-1297. arXiv:2506.09338.

Marco Saerens, Patrice Latinne, and Christine Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. *Neural Computation*, 14(1):21–41, 2002.

Ankur Samanta, Akshayaa Magesh, Ayush Jain, Kavosh Asadi, Youliang Yu, Daniel Jiang, Boris Vidolov, Kaveh Hassani, Paul Sajda, Jalaj Bhandari, and Yonathan Efroni. Structure enables effective self-localization of errors in LLMs. In *International Conference on Machine Learning*, 2026. arXiv:2602.02416.

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. 13449–13462, 2025. doi: 10.18653/v1/2025.emnlp-main.679.

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.

Lin Sun, Chuang Liu, Xiaofeng Ma, Tao Yang, Weijia Lu, and Ning Wu. FreePRM: Training process reward models without ground truth process labels. *arXiv preprint arXiv:2506.03570*, 2025.

sVanZeno. Forced disagreement: Operating points in consensus filtering for process reward models. GitHub repository sVanZeno/Operating-Points-in-Consensus-Filtering-for-Process-Reward-Models, 2026. Code release, accessed 2026-09-26.

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.

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/. arXiv:2312.08935.

Lingyan Wu, Xiang Zheng, Weiqi Zhai, Wei Wang, Xuan Ren, Zifan Zhang, Hu Wei, and Bing Zhao. MedPRMBench: A fine-grained benchmark for process reward models in medical reasoning. *arXiv preprint arXiv:2604.17282*, 2026.

Yuanhao Wu, Juntong Song, Hanning Zhang, Tong Zhang, and Cheng Niu. DuaShepherd: Integrating stepwise correctness and potential rewards for mathematical reasoning. *arXiv preprint arXiv:2506.17533*, 2025.

Bin Xie, Bingbing Xu, Xueyun Tian, Yilin Chen, and Huawei Shen. Towards robust process reward modeling via noise-aware learning. *arXiv preprint arXiv:2601.12748*, 2026.

Zhaohui Yang, Chenghua He, Xiaowen Shi, Linjing Li, Qiyue Yin, Shihong Deng, and Daxin Jiang. Beyond the first error: Process reward models for reflective mathematical reasoning. *arXiv preprint arXiv:2505.14391*, 2025.

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.

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

Lingyin Zhang, Jun Gao, Xiaoxue Ren, and Ziqiang Cao. The bidirectional process reward model. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12564–12580, 2026. doi: 10.18653/v1/2026.acl-long.572. URL https://aclanthology.org/2026.acl-long.572/. arXiv:2508.01682.

Yao Zhang, Yu Wu, Haowei Zhang, Weiguo Li, Haokun Chen, Jingpei Wu, Guohao Li, Zhen Han, and Volker Tresp. GroundedPRM: Tree-guided and fidelity-aware process reward modeling for step-level reasoning. *arXiv preprint arXiv:2510.14942*, 2025a.

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*, 2025b. arXiv:2501.07301.

Jian Zhao, Runze Liu, Kaiyan Zhang, Zhimu Zhou, Junqi Gao, Dong Li, Jiafei Lyu, Zhouyi Qian, Biqing Qi, Xiu Li, and Bowen Zhou. GenPRM: Scaling test-time compute of process reward models via generative reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 34932–34940, 2026. doi: 10.1609/aaai.v40i41.40797.

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/. arXiv:2412.06559.

Yefan Zhou, Yilun Zhou, Austin Xu, Soroush Vosoughi, Shafiq Joty, and Jiang Gui. The hidden signal of verifier strictness: Controlling and improving step-wise verification via selective latent steering. *arXiv preprint arXiv:2605.20745*, 2026.

Jiachen Zhu, Congmin Zheng, Jianghao Lin, Kounianhua Du, Ying Wen, Yong Yu, Jun Wang, and Weinan Zhang. Retrieval-augmented process reward model for generalizable mathematical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 8453–8468, 2025. doi: 10.18653/v1/2025.findings-acl.444.

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*, 2025. arXiv:2506.18896.

A Benchmark and Cohort Details

Table 6: Evaluation scope. Metrics have different meanings and are never averaged across benchmarks. Counts are the headline cohorts used throughout.

Benchmark	Records	Metric	Cohort note
ProcessBench	3,400	Macro-F1	All four mathematical subsets.
PRMBench	6,211	PRMScore	Non-clean construction records; 758 clean controls remain in the score cache but not the headline metric.
MedPRMBench	6,500	Step Macro-F1	Medical PRM benchmark (Wu et al., 2026); corrupted chains from the paired 13,000-chain evaluation set.
AgentProcessBench	999	First-error accuracy	Common no-truncation cohort; one official trajectory is excluded for supported context length.

ProcessBench reports a harmonic mean of exact first-error localization and clean-chain accuracy within each of GSM8K, MATH, OlympiadBench, and OmniMath, then averages the four values (Zheng et al., 2025). PRMBench and MedPRMBench instead evaluate pointwise step decisions and may contain multiple errors. Consequently, a threshold that maximizes one column need not maximize another even for perfectly preserved score ordering.

The frozen medical export has 6,500 unique cases, each with a clean and a corrupted chain for the same question. All clean step labels are positive and every corrupted chain has an error label. The text before the first error is identical in 5,098 pairs; the other 1,402 contain earlier wording changes, so "paired" means case-matched rather than a strict identical-prefix edit. Normalized exact-question overlap is zero against the 6,144-chain fixed-support training file, its 5,002-case source validation file, and the frozen ProcessBench/PRMBench questions; the 6,879-case medical source training split also has zero matching case IDs or normalized questions. This exact-match check does not establish semantic nonoverlap or independent label quality.

AgentProcessBench labels assistant messages as positive, neutral, or negative and separately reports first-error accuracy (Fan et al., 2026). For this task neutral is treated as nonnegative, and a prediction is the exact original message index of the first negative step or "no error." The common subset contains 250 trajectories each from BFCL, HotpotQA, and τ^2-Bench and 249 from GAIA. Always predicting no error obtains 38.64%; only the two RLHFlow checkpoints cover this cohort under the audited native adapters. Shorter-context models are not evaluated on easier model-specific subsets.

The exact breakpoint audit of these 21 model–benchmark cells raises the previous 0.001-grid Best by at most 0.027 metric points. The table and stars use exact maxima; the drawn curves retain the 0.001 plotting grid.

B Native-Interface Audit

Standard scalar and binary-token boundaries. The comparable matrix contains Qwen2.5-Math-PRM-7B, Qwen2.5-Math-7B-PRM800K, Math-Shepherd-7B, RLHFlow-Mistral-8B, RLHFlow-DeepSeek-8B, Universal-PRM-7B, and ReasonFlux-PRM-7B. Each exposes a probability score or equivalent binary logit margin, so 0.5 defines a common zero-margin convention. This does not imply that every model's authors recommend first-crossing deployment at that point: where no explicit binary promise is documented, "native" in the diagnostic denotes the origin induced by the released emission, not an official tuned threshold. The matrix has three complete benchmark cells per model plus the two supported AgentProcessBench cells, for 23 cells total. Missing cells are retained as missing rather than filled from a different context policy.

The emitted-score adapters for the Qwen-Math, Math-Shepherd, RLHFlow, and Universal-PRM rows and the ReasonFlux interface, with a copy of its model card, are included in the supplementary material. These records verify the scoring interface, not an author-approved binary deployment threshold.

Table 7: **Three-benchmark reference summary, including MedPRMBench.** Columns are ProcessBench four-subset first-error Macro-F1, PRMBench PRMScore, and MedPRMBench step Macro-F1 (%). Public rows are complete local replays: Skywork uses its released reward-drop rule, while ReasonEval and Llemma use released score emissions at a diagnostic 0.5 boundary. Other public 0.5 readers are also diagnostic conventions. Ours are 20K-chain pooled means $\pm$ sample SD over seeds 42, 77, and 123. Rows are not matched in data, backbone, or interface; no row is a proposed method. Interface details and cohort breakdowns appear in the appendix. Bold: column best; underline: best released PRM.

PRM	ProcessBench ↑	PRMBench ↑	MedPRMBench ↑
Released PRMs (complete local replay)			
Qwen2.5-Math-PRM-7B	<u>73.1</u>	<u>65.5</u>	62.6
Qwen2.5-Math-7B-PRM800K	56.4	58.3	43.3
Math-Shepherd-7B	28.6	47.6	55.6
RLHFlow-Mistral-8B	28.6	54.4	42.5
RLHFlow-DeepSeek-8B	26.8	53.9	49.1
ReasonFlux-PRM-7B	44.8	55.2	39.2
ReasonEval-7B	35.0	59.8	<u>70.4</u>
Llemma-PRM800K-7B	37.8	62.0	60.0
Skywork-PRM-1.5B	32.0	61.5	45.7
Skywork-PRM-7B	32.3	61.4	50.8
Universal-PRM-7B	0.0	45.7	38.5
Controlled testbed (ours), pooled scalar PRMs			
Qwen3.5-4B pooled (ours)	69.3 ± 0.5	67.4 ± 0.1	69.0 ± 1.2
Qwen3.5-9B pooled (ours)	$\mathbf{75.2} \pm 0.5$	$\mathbf{68.5} \pm 0.1$	$\mathbf{70.9} \pm 0.6$

Table 8: **MedPRMBench taxonomy-resolved fixed-reader context.** PRMScore, Acc., and First. are respectively corrupted-chain step macro-F1, step accuracy, and exact first-error accuracy. The 14 code columns are error-step recall, attributed by the benchmark's `error_step_code_map`; multi-coded error steps contribute to each of their codes. Five corrupted chains without a released code map remain in the overall metrics but not code columns. Probability heads use 0.5; Skywork uses its released cumulative reward-drop rule. Public rows are complete 6,500-chain local replays. For our rows, PRMScore is the stated multi-seed mean $\pm$ sample SD; all diagnostic cells are the corresponding seed means.

	Overall			Simplicity		Soundness							Sensitivity				
PRM	PRMScore	Acc.	First.	NR.	NCL.	EBS.	SC.	CA.	CI.	SA.	IGC.	TR.	PS.	DR.	MSC.	QC.	DDC.
Public checkpoints (local replay)																	
Qwen2.5-Math-PRM-7B	62.6	71.0	37.0	2.0	32.8	34.9	43.1	38.6	59.4	30.0	50.4	35.2	35.8	54.3	41.7	71.2	47.3
Qwen2.5-Math-7B-PRM800K	43.3	64.1	7.8	0.2	4.7	5.1	7.5	5.7	11.4	2.7	6.0	5.0	7.1	2.2	1.2	28.7	8.5
Math-Shepherd-7B	55.6	55.6	28.4	77.0	71.2	74.0	80.2	72.2	78.4	69.9	74.3	76.9	63.2	72.7	60.1	90.4	55.5
RLHFlow-Mistral-8B	42.5	61.3	16.7	0.0	6.7	5.2	3.9	6.2	11.6	5.5	7.2	5.6	6.0	6.4	9.4	9.9	11.3
RLHFlow-DeepSeek-8B	49.1	64.5	22.3	0.6	12.8	14.3	15.9	17.6	27.9	14.3	21.7	15.3	14.3	25.0	20.5	28.2	24.4
Skywork-PRM-1.5B	45.7	61.9	9.4	9.8	10.1	9.0	12.6	9.6	9.2	7.4	11.3	8.6	7.8	13.2	4.1	24.0	6.7
Skywork-PRM-7B	50.8	64.8	12.9	12.6	15.0	15.3	22.3	16.9	20.4	13.7	17.0	16.3	15.2	24.5	12.9	21.7	12.7
Universal-PRM-7B	38.5	62.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Ours: PRM800K, unweighted pooled-step BCE																	
Ours Qwen3.5-4B pooled ($n = 3$)	$69.0_{\pm 1.2}$	75.0	50.2	2.5	29.5	46.5	62.0	48.6	76.0	41.9	52.1	49.2	41.0	53.9	65.8	84.8	50.7
Ours Qwen3.5-9B pooled ($n = 3$)	$70.9_{\pm 0.6}$	76.3	59.1	2.9	29.6	48.5	59.2	50.8	78.9	44.1	64.4	51.5	41.0	74.0	70.6	92.5	52.7

For each frozen score file, the audit requires finite scores, exact alignment with step labels, the expected cohort cardinality, and parity with the reference evaluator at thresholds 0, 0.1, 0.5, 0.8, 0.9, and 1. The dense sweep is admitted only after all six checks pass. Across scalar models, the same benchmark-specific reader uses a shared threshold: first crossing on ProcessBench and pointwise step decisions on PRMBench and MedPRMBench. This standardizes the score-to-decision map within each benchmark without changing model emissions.

PRMBench abbreviations in table 2b are NR. (redundancy), NCL. (non-circularity), ES. (counterfactual error), SC. (step contradiction), DC. (domain inconsistency), CI. (confidence), PS. (missing

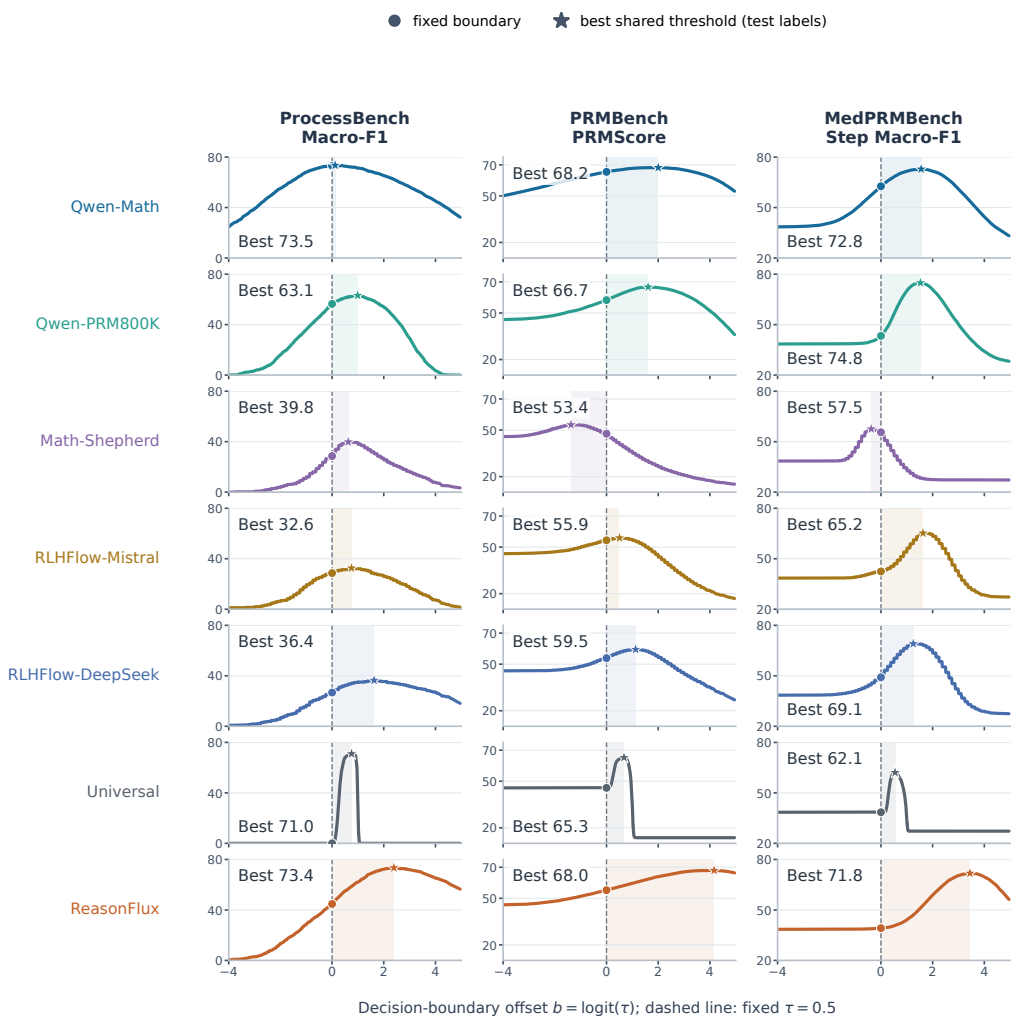


Figure 4: **Threshold-scan curves for seven released scalar PRMs across three benchmarks.** Each panel holds one model's frozen step scores fixed and moves only a shared score boundary. ProcessBench uses first crossing, as in fig. 1a,b; PRMBench and MedPRMBench threshold each step. The dashed line and circle mark the emitted-score 0.5 convention; the star marks the exact observed-score breakpoint maximizing the metric with target test labels. Shading shows the signed boundary displacement. Each panel reports the star's Best score in that benchmark's metric points. A test-label scan maximum is diagnostic, not a deployable threshold; 0.5 is not necessarily an author-endorsed binary rule. Each benchmark has its own metric and vertical scale. Skywork's cumulative reward-drop and generative PRMs use different interfaces and are reported separately.

condition), DR. (deception), and MS. (multiple solutions). The detailed-table evidence records the SHA-256 digest of every source file.

Cumulative-value interface. Skywork emits cumulative values rather than step-correctness probabilities. Its released rule flags an error when the reward drop crosses -0.05. We scan that reward delta directly instead of relabeling probability 0.5 as native. The resulting radii are meaningful within a Skywork row but cannot be compared numerically with logit radii.

Table 9: **Public PRMs: fixed and recoverable decision performance.** Fixed is the emitted-score zero-margin convention; Best is the exact observed-score breakpoint maximum using test labels and is diagnostic only. Gap is Best minus Fixed. Public 0.5 is not an endorsed binary rule (table 10). AUROC is observed-prefix step AUROC; on ProcessBench it uses only the annotated prefix through the first error; post-error suffixes have no step-correctness label. Scores are comparable only within a benchmark.

	ProcessBench (Macro-F1)				PRMBench (PRMScore)			
PRM	Fixed↑	Best↑	Gap↓	AUROC↑	Fixed↑	Best↑	Gap↓	AUROC↑
Qwen2.5-Math-PRM-7B	73.10	73.50	0.40	0.954	65.47	68.23	2.76	0.798
Qwen2.5-Math-7B-PRM800K	56.40	63.11	6.72	0.915	58.30	66.71	8.41	0.795
Math-Shepherd-7B	28.59	39.76	11.17	0.729	47.63	53.37	5.73	0.586
RLHFlow-Mistral-8B	28.61	32.58	3.97	0.691	54.38	55.92	1.54	0.626
RLHFlow-DeepSeek-8B	26.76	36.40	9.65	0.726	53.89	59.49	5.60	0.690
Universal-PRM-7B	0.00	70.97	70.97	0.883	45.72	65.28	19.56	0.725
ReasonFlux-PRM-7B	44.76	73.36	28.60	0.953	55.24	67.96	12.72	0.789



Figure 5: **The two released Skywork PRMs in their native reward-drop interface.** Each curve varies the threshold on consecutive cumulative-reward differences while holding scores fixed. The dashed line and circle mark the released -0.05 rule; the star is the test-label best shared threshold. Each panel reports its Best score in metric points. This test-label optimum is diagnostic, not a deployable result. The shaded horizontal interval shows its displacement from the released boundary. Horizontal units are reward deltas and are not comparable with the probability-logit offsets in fig. 4.

Three-class validity interface. ReasonEval-7B is replayed with its released negative/neutral/positive head and step score $P(\text{neutral}) + P(\text{positive})$. At the stated diagnostic 0.5 convention, it obtains 35.04 ProcessBench Macro-F1, 59.76 PRMBench PRMScore, and 70.44 MedPRMBench corrupted-chain step Macro-F1. Exact test-label breakpoint maxima are 44.07, 65.28, and 74.17, respectively; these are diagnostic ceilings rather than deployable calibrations. Its medical corrupted-chain first-error accuracy is 47.40%, and clean-chain acceptance is 47.08%. All 23,369 chains and 233,810 steps pass shard-hash coverage and six-threshold reference-evaluator parity checks.

Mean-vocabulary-logit interface. Llemma-PRM800K-7B uses its model card's $\text{sigmoid}(\text{mean}_v\, z_v)$ readout on an inserted double newline after each step. Tokenizer offsets select only these step boundaries, not newlines already inside a step; the card's own example is reproduced to within 0.003 probability per step. At a diagnostic 0.5 boundary, the full replay obtains 37.83 ProcessBench Macro-F1, 61.97 PRMBench PRMScore, and 59.98 MedPRMBench corrupted-chain step Macro-F1. Its exact test-label maxima are 41.87, 65.67, and 63.40, respectively. An initial synthetic-EOS variant assigned an almost constant 0.42 score to the final step and is retained as a superseded method check, outside the paper's comparisons. The corrected newline

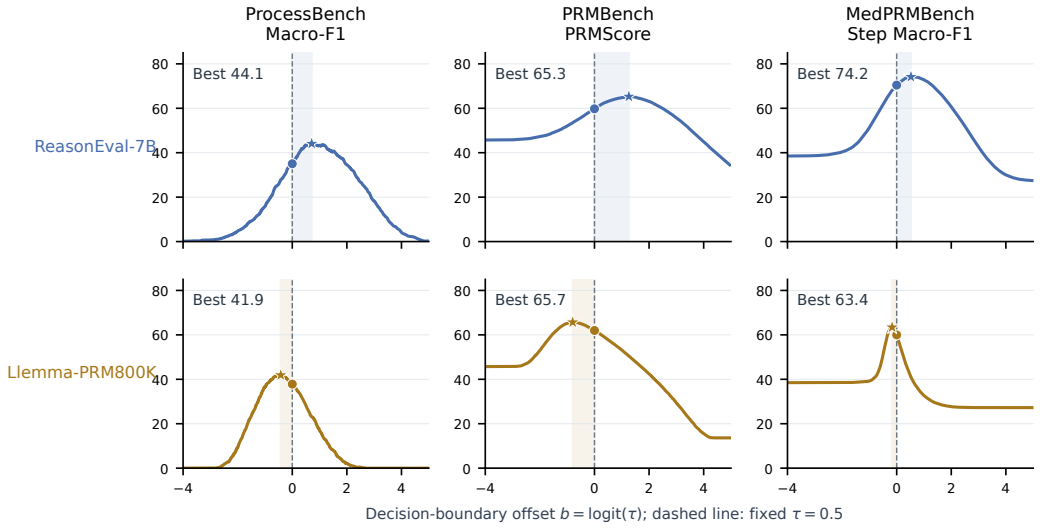


Figure 6: **Additional public validity readouts.** ReasonEval uses $P(\text{neutral}) + P(\text{positive})$; Llemma uses the released mean-vocabulary-logit readout at inserted step newlines. Curves hold each score cache fixed and vary a single threshold. Circles mark diagnostic 0.5 readers; stars mark exact best shared thresholds selected with test labels, which are not deployable calibrations. Metrics differ by benchmark.

Table 10: **Provenance of the seven public zero-margin conventions.** Score emissions follow the audited adapters. "Not established" means the documentation inspected for this paper does not establish an author-endorsed 0.5 first-crossing deployment rule; it is not a claim that no such rule exists elsewhere. All seven use the same diagnostic reader in table 9.

Model	Emitted-score basis for 0.5	Binary-rule status
Qwen2.5-Math-PRM-7B	Positive class of two-class process head	Not established
Qwen2.5-Math-7B-PRM800K	Positive class of two-class process head	Not established
Math-Shepherd-7B	Normalized $+/-$ token scores	Not established
RLHFlow-Mistral-8B	Normalized $+/-$ token scores	Not established
RLHFlow-DeepSeek-8B	Normalized $+/-$ token scores	Not established
Universal-PRM-7B	Sigmoid scalar prefix score	Not established
ReasonFlux-PRM-7B	Positive class of two-class process head	Not established

reader passes all 23,369 chains, 233,810 steps, shard hashes, and six-threshold reference parity checks.

Generative verdict interface. R-PRM-7B is evaluated through its documented critique–verdict path with one sampled critique ($N = 1$). The final positive–negative verdict-logit margin has a native zero boundary. It obtains 63.86 Fixed and 64.01 Best ProcessBench Macro-F1, a gap of 0.15 points, and Radius@1 of 1.11 logits. This establishes that the Fixed–Best analysis applies to a binary verdict after generated reasoning; it is not an official reproduction of the paper's larger voting budget. The same $N = 1$ replay obtains 65.55 PRMBench PRMScore and 54.67 MedPRMBench corrupted-chain step Macro-F1 at a fixed Yes-probability boundary of 0.5. On the medical pairs, corrupted-chain first-error accuracy is 33.25% and clean-chain acceptance is 1.28%. These measurements use one sampled critique per step and therefore have a different inference budget from scalar PRMs; five PRMBench chains required the predefined context-trimming rule.

EurusPRM is excluded from the standard-boundary matrix because its documented ProcessBench boundary is source-selected rather than a native 0.5 rule. Unofficial direct-verdict probes of gen-

Table 11: **Skywork Fixed, Best, and gap in the released cumulative-reward-delta space.** Fixed uses the documented -0.05 reward-drop threshold. The best threshold $\delta^\star$ and Radius@1 are in reward-delta units and are not numerically comparable with the logit quantities in fig. 4.

PRM	Benchmark	Fixed	Best	Gap	$\delta^\star$	Radius@1
Skywork-7B	ProcessBench	32.31	37.69	5.37	-0.121	0.005
	PRMBench	61.40	63.07	1.66	-0.095	0.010
	MedPRMBench	50.81	55.65	4.84	-0.019	0.005
Skywork-1.5B	ProcessBench	32.00	35.84	3.84	-0.132	0.005
	PRMBench	61.53	62.76	1.23	-0.083	0.010
	MedPRMBench	45.69	51.04	5.35	-0.012	0.006

erative systems are kept outside the formal comparison when their released pipeline or execution semantics cannot be reproduced.

C Fixed-Support Paired Effects and 9B Replication

Table 12: **Every fixed-support paired effect on ProcessBench.** All changes are from the same-seed, same-reduction ratio-one baseline. Mean/Pooled denote chain-mean/pooled-step reduction. Best is a diagnostic test-label score; a smaller gap need not raise Fixed. Forward 4B Fixed scores improve in 11/12 pairs; forward 9B Fixed scores improve in 2/3. Event AUC contrasts the first-error position with every other position and is not step-correctness AUROC.

Scale	Intervention, seed	ΔFixed	ΔBest	ΔGap	ΔEvent AUC
4B	30% Mean, s42	−0.57	−2.95	−2.38	−0.0253
4B	30% Mean, s77	+4.66	+2.48	−2.17	+0.0124
4B	30% Mean, s123	+2.92	+0.93	−1.99	+0.0118
4B	30% Pooled, s42	+3.96	+0.08	−3.88	−0.0070
4B	30% Pooled, s77	+0.42	−0.88	−1.30	−0.0161
4B	30% Pooled, s123	+6.51	+1.14	−5.36	+0.0076
4B	15% Mean, s42	+2.29	−0.44	−2.73	−0.0061
4B	15% Mean, s77	+2.00	−0.10	−2.10	−0.0086
4B	15% Mean, s123	+2.79	+0.73	−2.06	+0.0076
4B	15% Pooled, s42	+6.14	+0.69	−5.44	+0.0065
4B	15% Pooled, s77	+2.11	−1.31	−3.42	−0.0163
4B	15% Pooled, s123	+8.96	+3.21	−5.74	+0.0171
4B	Reverse Mean, s42	−6.79	−0.25	+6.54	−0.0022
4B	Reverse Pooled, s42	−6.99	+0.71	+7.70	−0.0048
9B	15% Mean, s42	+2.98	+0.48	−2.50	−0.0019
9B	15% Mean, s77	−1.03	−1.92	−0.89	−0.0081
9B	15% Mean, s123	+0.59	+0.05	−0.54	−0.0065
9B	Reverse Mean, s42	−4.76	−0.56	+4.20	+0.0005

The target-prior minus baseline paired-bootstrap change in the gap is -1.48 points (95% CI $[-2.92, -0.33]$); the Best-threshold logit shift is -0.652 (95% CI $[-1.017, -0.279]$). The first-error-versus-rest AUROC changes by -0.0055 (95% CI $[-0.0082, -0.0022]$). This diagnostic treats the unlabelled post-error suffix as non-first; it cannot establish a change in step-correctness ranking. Its Fixed score still falls for one seed (77); the bootstrap CI applies to the gap, not to an asserted improvement in Fixed Macro-F1.

C.1 Second-Backbone Pilot: Llama-3.1-8B

We also ran a pre-specified single-seed, three-arm fixed-support intervention with Meta-Llama-3.1-8B. The same 6,144 training chains, 5,002 source-validation chains, example order, training budget, and seed 42 are shared across arms; only the incorrect-to-correct loss-weight ratio changes (table 14). The base tokenizer needed one common local chat-template adapter, recorded with hashes

Table 13: **Per-seed Qwen3.5-9B fixed-support replication.** Baseline and 15%-prior rows are paired within seed; the reverse-direction control was pre-specified for seed 42 only. Best and Radius@1 use the frozen ProcessBench evaluation labels diagnostically. Event AUC contrasts the first error with every other position, including the unlabelled suffix; it is not step-correctness AUROC.

Role	Seed	w_-/w_+	Event AUC	Fixed	Best	Gap	τ_{src}	R@1
Baseline	42	1.000	0.8869	68.33	71.06	2.73	0.585	0.144
15% prior	42	1.925	0.8850	71.32	71.55	0.23	0.440	0.352
Baseline	77	1.000	0.8930	70.40	71.62	1.22	0.550	0.319
15% prior	77	1.925	0.8848	69.37	69.70	0.33	0.370	0.405
Baseline	123	1.000	0.8921	70.94	71.79	0.85	0.615	0.156
15% prior	123	1.925	0.8856	71.53	71.84	0.31	0.375	0.229
Reverse 0.5	42	0.500	0.8874	63.57	70.51	6.93	0.790	0.104

in the supplementary material. All arms were evaluated on the same 3,400 ProcessBench chains in the same order.

Table 14: **A single-seed Llama-3.1-8B pilot limits cross-backbone claims.** Best is the exact saved-score, test-label shared-threshold maximum and is only diagnostic. Prefix AUROC excludes unlabelled post-error suffixes. First-error and clean columns are four-subset macro accuracies at 0.5.

w_-/w_+	Fixed	Best	Gap	Prefix AUROC	First error	Clean accept
1.000	24.66	29.76	5.10	0.7350	15.98	74.75
1.925	20.68	21.03	0.35	0.6563	14.69	54.57
0.500	26.19	35.37	9.18	0.7853	16.24	84.16

The forward intervention moves ProcessBench step logits by median -0.797 against a weighted-BCE prediction of -0.655, but the annotated correct steps move more than first-error steps (-1.109 versus -0.488). Its fixed Macro-F1 falls by 3.98 points (95% paired chain-bootstrap interval $[-6.46, -1.68]$), and Best falls by 8.73 points. The smaller gap therefore does not demonstrate a useful fixed reader. The baseline's source-selected threshold scores 29.08, versus 20.68 for the forward arm at 0.5; neither threshold was selected on ProcessBench labels. The reverse arm moves logits by median $+0.672$ against a prediction of $+0.693$, while its Best and Prefix AUROC both increase. Translating baseline decision curves by the predicted constant improves forward curve RMSE only from 8.79 to 8.12 points and *worsens* reverse RMSE from 4.73 to 7.60. Thus the signed movement transfers, but class separation and curve shape also change. One seed and a weak baseline cannot establish a family-level negative result; the experiment does not support an architecture-independent pure-shift claim. Full hashes, metrics, and conditional bootstrap intervals are in the supplementary material.

D Frozen Mechanism Diagnostics

We rerun the 12 forward ProcessBench interventions over every observed score breakpoint and its next representable floating-point value. The largest exact Best increase relative to the original 0.001 grid is 0.029 points (mean 0.014), so the grid does not affect the reported directions or their scale. For the curve-collapse check, we translate each paired baseline curve by $-\log(w_-/w_+)$ and compare it with the weighted model over $|b| \leq 4$ logits. The mean RMSE falls from 6.53 to 2.60 points. Step-level residuals remain: their mean within-pair shift standard deviation is 0.752 logits, even as the mean rank correlation is 0.948. Two identical training runs in the rebuilt-backbone rerun (section G) differ by as much (per-step shift SD 0.60–1.36 logits, rank correlation 0.85–0.96; its weighted-versus-baseline pairs: 0.63–1.27 and 0.86–0.96), so these residuals are at the level of training nondeterminism. The full per-intervention quantiles, exact optima, and source-transfer cells are in the supplementary material.

The few-shot replay uses complete frozen target caches and record-disjoint remainders. PRM-Bench thresholds optimize official PRMScore on its 6,211 non-correct records; MedPRMBench

Table 15: **Few-shot target calibration on frozen scores.** One global threshold is selected on k labelled target records and evaluated on the disjoint remainder. Prior only uses the same k records only to estimate the error rate $\hat{\pi}$ and flags that fraction of the unlabelled remainder. Values average 12 intervention models and 100 deterministic category-stratified splits. Gaps to the full-target Best are leaky diagnostic upper bounds, not deployment results.

Benchmark	k	Zero-margin	Source	Few-shot	Prior only	Gap to Best
PRMBench	16	66.38	66.59	65.38	66.43	1.34
	64	66.37	66.58	66.21	66.58	0.50
	400	66.38	66.58	66.53	66.63	0.18
MedPRMBench	16	65.19	68.57	72.01	72.52	1.02
	64	65.18	68.57	72.57	72.88	0.46
	400	65.18	68.57	72.87	72.94	0.16

uses corrupted-chain step Macro-F1. Split seeds, threshold variance, intervals, and every model–split cell are in the supplementary material. The Prior-only column asks what the labelled records contribute. Estimating only the error rate from them and flagging that fraction of the remainder beats the selected threshold for every model at every k except one model at $k = 400$ on MedPRMBench, is on average within 0.44 points of the same reader given the remainder's true error rate, and raises the tenth percentile over splits at $k = 16$ (66.0 vs. 63.1 on PRMBench, 71.7 vs. 70.4 on MedPRMBench, averaged over models). On these targets the labels mostly supply one scalar, the error rate. This is a diagnostic of what the labels carry, not a proposed reader.

A fourth, agentic target. Table 16 replays the two unweighted 4B reductions (chain mean and pooled steps, 20K PRM800K chains) on the same 999 AgentProcessBench trajectories at seeds 42/77/123, scoring first-error accuracy. At the fixed 0.5 reader pooled steps lead by 12.81, 20.42, and 3.60 points. When *each* recipe is read at its own test-label optimum, the lead shrinks to 3.80, 6.41, and -0.50 points: on average a single shared threshold removes about three quarters of the fixed-reader gap (mean $12.28 \rightarrow 3.23$), and at seed 123 it removes all of it. The remainder is seed-dependent and is what a shared threshold cannot recover; it can reflect ranking, subgroup-specific score scales, or other finetuning changes, and is not an identified representation effect. Always predicting "no error" scores 38.64%, so fixed-reader agent accuracy is also sensitive to the benchmark's label prior.

Table 16: **How much of the agent gap can one threshold recover?** First-error accuracy of the two unweighted 4B reductions on the same 999 chains. Best uses target labels and is not deployable. The last column compares both recipes at their own Best threshold, not one calibrated model against the other model's default.

Seed	Unweighted reduction	Fixed 0.5	Best	$\tau^\star$	Pooled$-$mean
42	Chain mean	35.94	46.15	0.190	–
42	Pooled steps	48.75	49.95	0.376	+3.80
77	Chain mean	23.02	38.74	0.096	–
77	Pooled steps	43.44	45.15	0.356	+6.41
123	Chain mean	40.04	45.25	0.201	–
123	Pooled steps	43.64	44.74	0.481	-0.50

E THRESHOLDED FIXED-POOL CHECK

The MATH-500 experiment changes a binary step gate on a fixed pool of 64 candidates per problem (table 17); it is not a policy-optimization experiment. The accuracy intervals below cross zero, so the result bounds the downstream claim to threshold sensitivity rather than establishing an accuracy gain.

Table 17: **Threshold-dependent MATH-500 fixed-pool selection.** Accuracy is in percent, averaged over three paired 9B seeds; each B/T entry is baseline/15%-prior. Continuous uses minimum step score. Fixed and Source use the mean fraction of steps passing 0.5 or the frozen source-validation threshold. ΔFixed is target minus baseline; ΔSourceGain is the target-minus-baseline change in (Source−Fixed). Intervals are 95% hierarchical seed/problem bootstrap CIs.

N	Continuous B/T	Fixed B/T	Source B/T	ΔFixed	ΔSourceGain
1	65.40/65.40	65.40/65.40	65.40/65.40	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]
4	70.07/69.93	68.13/68.47	68.73/68.20	+0.36 [−0.67, +1.40]	−0.87 [−1.93, +0.13]
8	72.67/72.40	68.67/69.33	69.80/68.87	+0.65 [−0.53, +1.80]	−1.58 [−2.80, -0.47]
16	73.47/72.73	68.73/69.33	69.73/69.13	+0.60 [−0.73, +2.13]	−1.19 [−2.47, +0.00]
32	72.33/72.07	68.93/69.20	69.80/69.13	+0.26 [−0.93, +1.47]	−0.92 [−2.00, +0.20]
64	71.13/71.27	68.87/69.13	69.73/69.13	+0.29 [−0.80, +1.47]	−0.87 [−2.00, +0.27]

F Interpretation Boundaries

Fixed, Best, and the gap are defined for a model–benchmark–metric triple. In particular, the optimal threshold of a ratio metric depends on the evaluation distribution; even calibrated probabilities need not have an F1-optimal threshold of 0.5 (Lipton et al., 2014). Cross-benchmark threshold movement therefore does not by itself identify score miscalibration or a training defect. They answer a narrower operational question: given the released scores and native interface, how much benchmark performance is available to one shared boundary, how much the specified zero-margin boundary discards, and how locally sensitive that decision is.

The three reported quantities cannot replace one another. Best without the gap hides deployment mismatch. The gap without Best can make a weak model appear healthy. Radius without both can reward a broad, uniformly poor plateau, as Universal-PRM demonstrates on ProcessBench. Best thresholds use target labels and are descriptive scan maxima, never proposed calibration for a new domain.

Finally, this evaluation does not establish downstream reward hacking. If a system consumes a thresholded gate, changing the boundary changes its reward directly. If it ranks candidates by a continuous score, a monotone score shift may leave ranking unchanged. Demonstrating reward hacking additionally requires optimization pressure under which proxy reward rises while an independent true objective deteriorates; no such policy-optimization claim is made here.

G Prevalence-Quantile Diagnostic: Sensitivity and Follow-Ups

All numbers below are exploratory. The public score caches were inspected before the diagnostic was proposed, and π is a target-label quantity. Scripts and outputs for every analysis below, including the fixed-support re-run, are in the supplementary material.

How optimistic is Best? Best chooses its threshold on the evaluation labels. Choosing it instead on one half of the chains (stratified by subset on ProcessBench) and scoring the other half, averaged over both folds and 60 random splits, lowers Best by 0.22 points on average over the 21 public cells and by at most 0.64 (Math-Shepherd on ProcessBench). The gaps in section 4 are therefore not an artefact of in-sample selection.

Misspecified prevalence. Across the 21 public cells, the mean shortfall from Best is 15.73 points for Fixed, 0.88 for Q_{π}, 5.53 for $Q_{0.7\pi}$, and 5.42 for $Q_{1.3\pi}$. The penalty for misspecification is concentrated on ProcessBench, where chain-level first crossing makes the curve steep around its peak; on PRMBench and MedPRMBench, $\pm 30\%$ costs at most a few points in every cell. A label-free estimate of π, the weight of the lower component of a two-component Gaussian mixture on logit scores, gives a mean shortfall of 4.07 points. It is near Best for strongly discriminating models and fails for weak ones (e.g., Math-Shepherd on PRMBench: 35.0 against a Best of 53.4), so we make no label-free claim.

Few labels: estimate a prior or tune a threshold? We draw k labelled chains uniformly, 200 times per k, and evaluate on the disjoint complement. We compare reading at the quantile of the estimated prevalence $\hat{\pi}$ against tuning the threshold directly on the same k chains. On the two step-level benchmarks, the prevalence estimate has the higher mean in $13/14$ cells at $k{=}8$ and $12/14$ at $k{=}16$. Its tenth percentile is higher on average (62.9 against 59.3 at $k{=}8$), and the two coincide by $k{=}128$. On ProcessBench, $\hat{\pi}$ from very few chains is noisy: at $k{=}8$ it wins in only $3/7$ models, and it hurts the already-aligned Qwen-Math. Averaged over models it is ahead at $k{=}16$ and 32 (49.9 vs. 38.7; 51.6 vs. 45.7), mainly because threshold tuning on few chains fails to leave the 0.5 point for heavily displaced models (Universal-PRM: 64.7 vs. 8.8 at $k{=}16$). Estimating one scalar is thus often a cheaper route to the operating point than searching for a threshold.

Why the error rate? A reader that flags exactly a π fraction predicts as many errors as the target contains, and therefore as many correct units; precision equals recall for both classes at once, which is the break-even point of Macro-F1. Under an equal-variance binormal score model ($n{=}200$k, $\pi \in \{0.16, 0.37, 0.65\}$, $d' \in \{0.5, 1, 1.5, 2.5\}$) the Macro-F1-optimal flag rate stays within 0.97–$1.19\,\pi$ and Q_π loses at most 0.14 points. On the public caches the optimal flag rate is 0.94–$1.22\,\pi$ on ProcessBench and 0.99–$1.43\,\pi$ on the step benchmarks. The optimum thus sits at or slightly above Q_π. On PRMBench the excess comes from its labels: steps after the first error that differ from the clean solution but are labelled correct make up 29% of correct steps, yet 62–70% of the false positives at Best for Qwen-Math, Qwen-PRM800K, Universal-PRM, and ReasonFlux (false-positive rate 0.28–0.31, against 0.04–0.06 before the first error); under first-error labels the median optimal flag rate is 0.99π. On MedPRMBench the excess remains under first-error labels (1.23π). The break-even argument is specific to step-level Macro-F1. ProcessBench's Macro-F1 is instead a subset average of harmonic means of two class-conditional accuracies and does not depend on prevalence at all, so its 0.94–$1.22\,\pi$ is an empirical regularity rather than break-even. On pooled steps, the median optimal flag rate over the seven PRMs is 1.21π on PRMBench and 1.11π on MedPRMBench for Macro-F1, but 2.33π and 1.35π for balanced accuracy, 2.01π and 1.37π for MCC, and 1.80π and 1.61π for error-class F1; another metric calls for another quantile (Lipton et al., 2014).

The operating point is stable in rank, not in score. Over 100 chain-level bootstrap resamples (best threshold on a 500-quantile grid), its standard deviation is 0.014–0.214 on the logit scale but 0.006–0.032 in rank (flag rate); for Qwen-Math on PRMBench it is 0.214 logits against 0.014 in rank. A reported threshold value is therefore not a transferable property of a PRM, whereas its flag rate relative to π is. Carrying a PRM's best logit threshold to another benchmark leaves a mean shortfall from Best of 5.46 points over the 42 source–target pairs (2.57 into PRMBench, 6.57 into ProcessBench, 7.24 into MedPRMBench), between Fixed (15.73) and Q_π with the target π (0.88). Within ProcessBench, π rises with difficulty (GSM8K 0.52, MATH 0.59, OlympiadBench 0.66, Omni-MATH 0.76) and the per-subset optimal flag rate tracks it (medians 0.98–$1.06\,\pi$), while the per-subset optimal logit drifts down (median 0.6 logits from GSM8K to Omni-MATH, six of seven PRMs) as the longer, harder chains lower clean-chain minima and raise first-error scores. One shared threshold is still within 0.8 points of per-subset optima; what breaks a fixed reader is the shift between benchmarks.

The medical transfer residual is a prevalence shift. For the 12 frozen 4B interventions of section 6, the step error rate of the evaluated cohorts is 0.158 on PRMBench and 0.374 on MedPRMBench. The source threshold flags 16% and 20% of their steps and recalls 0.44 and 0.43 of their errors, so it behaves alike on both; relative to the error rate it flags 1.02π on PRMBench (range 0.85–1.17) but 0.54π on MedPRMBench (0.41–0.69). We split the move of the best boundary from PRMBench to MedPRMBench, $+1.25$ logits on average (s.d. 0.37), into a prevalence part, obtained by reweighting one benchmark's error and correct steps to the other's error rate with its class-conditional scores unchanged, and a class-conditional remainder, averaging both orders. The prevalence part is $+1.17$ and positive for all 12 models; the remainder is $+0.08$ and negative for four. The closed-form label-shift value $\operatorname{logit}\pi_{\mathrm{Med}} - \operatorname{logit}\pi_{\mathrm{PRMB}} = 1.16$ (Saerens et al., 2002; Lipton et al., 2018) is close to the prevalence part, and Q_π on the same scores is within 0.07 points of Best on MedPRMBench (at most 0.20), against 4.46 for the source threshold. Under first-error labels, the semantics the 4B models were trained with, the error rates are 0.122 and 0.326, the move is

$+1.28 = +1.10 + 0.18$, and Q_π is within 0.09 points of Best. The residual thus mirrors the training prior of section 5 at test time: a change of class prior moves the log-odds boundary additively.

The source threshold is itself a prevalence quantile. Every 4B run selects it on the same locked validation set (5,002 chains, 84.9% with an error; observed-prefix step error rate $\pi_{\mathrm{val}} = 0.129$). On the 23 checkpoints of the rebuilt-backbone re-run described below (eight seed-42 arms, four seed-77 arms, and eleven repeats), its flag rate on validation steps is 0.81–$1.16\,\pi_{\mathrm{val}}$ (median $0.96\,\pi_{\mathrm{val}}$), and it lies within 0.20 logits (median) of the step Macro-F1 optimum there, although it maximizes a chain-level criterion that ignores prevalence; the coincidence is specific to this set's composition (re-weighted to 50% error chains, the same threshold would flag $1.20\,\pi_{\mathrm{val}}$). At 0.5, every checkpoint already under-flags on the locked validation set (a chain-level ρ of 0.46–0.90), below its ProcessBench ρ in all 23 (correlation 0.96), so the fixed rule's shortfall exists in-distribution rather than arising from the benchmark shift. PRMBench's error rate is close to π_{val}, so the threshold travels; MedPRMBench's is 2.9 times higher. Shifting the source threshold by $\operatorname{logit}\pi_T - \operatorname{logit}\pi_{\mathrm{val}}$, which needs only the target error rate, leaves 0.35 points below Best on MedPRMBench instead of 4.46 (better in all 12 cells; 0.69 instead of 2.24 under first-error labels), and 0.59 or 1.52 points if that rate is under- or overestimated by 30%. On PRMBench, where the shift is only 0.23 logits, it leaves 0.24 instead of 0.13. Because π_T comes from test labels, this is a check of the prevalence account, not a deployable reader.

Released PRMs: offsetting shifts. For the seven released PRMs the same split gives a move of only $+0.12$ logits, but as the sum of a prevalence part of $+0.59$ (positive for all seven) and a class-conditional part of -0.46 (negative for five). Qwen-Math, Qwen-PRM800K, and ReasonFlux, for instance, score correct medical steps 1.7–2.4 logits lower (median) than correct PRMBench steps. Under first-error labels the move is $+0.01 = +0.69 - 0.68$. The small benchmark share of $b^\star$ variance in section 4 therefore reflects offsetting effects, not benchmark invariance.

Chain length is a ranking limit. Under first crossing, the fraction of clean chains rejected at the shared optimum rises with chain length for weaker PRMs (Math-Shepherd $0.35 \to 0.76$, RLHFlow-Mistral $0.34 \to 0.93$, RLHFlow-DeepSeek $0.35 \to 0.90$ from 1–5 to 13+ steps) but not for Qwen-Math or ReasonFlux (about 0.10 throughout). A separate threshold per length bin recovers only 0.06–0.81 ProcessBench points, so this length tax sits on the discrimination side of our decomposition, not the operating-point side.

Where the weaker PRMs fail. On ProcessBench a shared threshold can locate an error only if the erroneous step scores below every earlier step. This holds for 85–88% of error chains under Qwen-Math, ReasonFlux, Qwen-PRM800K, and Universal-PRM and for 73% under Math-Shepherd, but for only 29% and 41% under RLHFlow-Mistral and RLHFlow-DeepSeek; even at their best thresholds, 67–69% of their error chains are flagged too early. Step AUROC is not a position artefact: averaged within step positions, it differs from the pooled value by -0.002 to $+0.062$. The first-step score already separates chains that later err from clean chains with AUROC 0.68–0.72 for Math-Shepherd and the RLHFlow models, against 0.55–0.62 for the other four, consistent with Monte-Carlo labels that estimate eventual success rather than step validity (Zhang et al., 2025b). For these three models, reading drops in score instead of levels (the better of two drop readers) raises PRMBench Best by 4.5–6.6 points but does not help on ProcessBench; we do not pursue such readers here.

What the readout changes in reported behaviour. Two conclusions often drawn from fixed-boundary scores change when the same scores are read at Q_π. First, PRMs are reported to favour positive rewards (Song et al., 2025). For the six PRMs with $\rho < 1$ on PRMBench, error-step recall at 0.5 is 0.00–0.31 and correct-step accuracy 0.91–1.00; at Q_π, recall rises to 0.22–0.46 while correct-step accuracy falls only to 0.87–0.90 (MedPRMBench: recall 0.00–$0.32 \to 0.49$–0.68, accuracy 0.94–$1.00 \to 0.72$–0.81). Part of the reported bias is thus an operating-point effect rather than a property of the ranking. Second, released PRMs seem to disagree about which units are wrong: the mean pairwise Cohen's κ of their decisions at 0.5 is 0.25, 0.17, and 0.09 on ProcessBench (chain minima), PRMBench, and MedPRMBench (steps). Read at each model's own Q_π, it rises to 0.53, 0.37, and 0.42, and it rises in 62 of the 63 pairs (44 of 45 without Universal-PRM, which flags nothing at 0.5); much of the disagreement is about how many units to flag. Accordingly, across the 21 cells the size of the gap follows the flag-rate mismatch $|\log\rho|$ (Spearman 0.76, ρ floored at 0.01) far more closely than AUROC (0.32).

Controlled testbed. On our own ProcessBench caches, pooled-step BCE runs without class weights have ρ between 0.94 and 1.09 at 4B and 9B, which is why Fixed is already near Best in those runs. More training data does not move this boundary: at the natural 15% clean-chain mixture, seed-42 pooled-step runs on the nested 1K, 3K, 10K, and 20K supports reach Fixed 59.4, 64.6, 68.1, and 68.8 and Best 60.2, 65.0, 68.7, and 69.3, while the source-validation threshold stays between 0.510 and 0.545. At a fixed 6,144 chains with the same loss, raising the clean fraction from 15% to 30% and 50% instead moves that threshold from 0.560–0.630 to 0.655–0.740 and 0.770–0.845 across three seeds, so the class prior, not the amount of data, sets where the boundary sits. Training on the same nested 1K/3K/10K supports with class-balanced weights ($w_-/w_+ \approx 6.8$) raises ρ to 1.53, 1.38, and 1.25, and Fixed falls to 0.3, 32.2, and 54.4, while Q_π stays within 1.2 points of Best ($22.2/23.0$, $57.2/58.3$, $66.8/66.9$). In the single-seed Llama-3.1-8B pilot, raising w_-/w_+ from 0.5 to 1 to 1.92 moves ρ from 0.56 to 0.68 to 0.98, with Q_π within 0.6 points of Best at each ratio. In both cases the loss prior moves the flag rate, and the rank position of the optimum does not move. The raw caches and the text-only 4B backbone files behind section 5 were not retained, so we re-ran its eight seed-42 arms on the same 50%-clean fixed support with a text-only backbone rebuilt from the released Qwen3.5-4B checkpoint. The re-run does not reproduce the originally reported scores exactly: Fixed is lower in all eight arms, by 0.1–6.6 points, and Best is lower by 1.4–7.0 points in five arms and within 0.4 points in the other three, whereas identical runs of one pooled-step arm already differ by up to 4.3 (Fixed) and 5.0 (Best) points (three or four runs per arm). Averaged over its identical repeats, each pooled-step arm except the reverse one comes within 0.2 points of its originally reported Best while its Fixed remains 0.8–2.1 points lower. The direction of the effect does reproduce. Across $w_-/w_+ = 0.5$, 1, and the two forward ratios, ρ rises from 0.54 to 0.70, 0.83, and 0.88 under pooled-step loss and from 0.70 to 0.84, 0.92, and 0.98 under chain-mean loss; the gap falls from 18.8 to 10.0 to 2.9–3.0 points and from 6.6 to 2.7 to 0.5; and Q_π stays within 1.1 points of Best in all eight arms. Best, however, moves by up to 1.8 points across the pooled-step arms and by 4.0 across the chain-mean arms. The repeats separate location from discrimination. Each pooled-step arm's curve location (the logit translation that best aligns its ProcessBench curve with the baseline's) varies by 0.14 logits across identical runs (within-arm SD) and moves by -0.87, -0.50, and $+0.79$ logits for the 15%- and 30%-prior forward arms and the reverse arm, against $-\log(w_-/w_+) = -0.83$, -0.50, and $+0.69$; Fixed moves with it (95% intervals exclude zero), whereas Best and step AUROC of the two forward arms stay within run-to-run variation of the baseline's (Best $+0.4$ and -0.2 points, AUROC $+0.006$ and $+0.002$; all $p \geq 0.5$); only the reverse arm's Best is lower, by 3.0 points (95% interval $[-5.7, -0.4]$), which its originally reported Best did not show (one-way ANOVA across the four arms: $p = 0.08$ for Best, 0.12 for AUROC). Scored on PRMBench (official PRMScore), the same checkpoints repeat this pattern. Curve location moves by -0.62, -0.35, and $+0.92$ logits, each 95% interval containing $-\log(w_-/w_+)$; Fixed moves by $+1.6$, $+1.3$, and -4.0 points (intervals exclude zero); the forward arms' Best differs from the baseline's by only -0.1 and $+0.2$ points (95% intervals ± 0.5; AUROC -0.002 and -0.001); and the reverse arm's Best and AUROC are again lower, by 0.6 points ($[-1.1, -0.1]$) and 0.009 ($[-0.017, -0.002]$). Decision by decision, a weighted run read at 0.5 makes the same ProcessBench prediction (first flagged step) as a baseline run read at logit $\log(w_-/w_+)$ on 85.6% of problems, as often as two identical runs agree at a common threshold (85.0%; 81.2% without the shift). At seed 77, the rebuilt pooled-step baseline lands at or above its originally reported scores in two identical runs (Fixed 66.3 and 65.0 vs. 64.1, Best 69.1 and 69.0 vs. 69.0) and the 15%-prior forward arm below them (65.8 vs. 66.2, 67.1 vs. 67.7); ρ again rises (0.86 to 0.91) and the gap again closes (3.4 to 1.3 points), but through a 2.0-point fall in Best rather than a rise in Fixed, within the variation between identical runs. The re-run thus supports the direction of the flag-rate and gap effects and resolves the loss prior's effect on location; an effect of the forward priors on Best, if any, is below its resolution (95% intervals of about ± 2.6 points per arm on ProcessBench and ± 0.5 on PRMBench).

H Exploratory Label-Free Target-Side Adaptation

This appendix records a post-hoc exploratory extension of the public-PRM audit. It does *not* convert ProcessBench into an unseen test: its score files, labels, and Fixed results were inspected before the feature and gates below were proposed. Its purpose is narrower: to test whether the frozen score geometry contains a plausible target-side signal for choosing whether to move a first-crossing boundary without placing target labels in the fitting path.

Score-only fitting protocol. For each completed chain c, define $m_c = \min_i p_{c,i}$, transform m_c to logit space, and fit a two-component Gaussian mixture to a calibration batch of these minima. The candidate boundary τ is the intersection of the two weighted component densities, mapped back through the sigmoid. No step label, chain label, or first-error index is passed to this fit. We then apply the ordinary shared first-crossing reader, "first i with $p_{c,i} < \tau$." This feature deliberately matches the ProcessBench reader; it is not claimed to be a general feature for pointwise metrics.

For the main deployment-shaped replay, three outer seeds reserve 100 score-only chains from each of ProcessBench's four public subsets (400 chains total), freeze the fitted boundary, and evaluate only on the remaining 3,000 chains. The latter use labels exclusively for the metric. Within every outer split, reported 95% intervals resample these test chains 1,000 times within (ProcessBench subset, clean/error) cells while keeping the calibration batch, mixture, and boundary fixed. We additionally repeat the score-only draw 128 times at batch sizes 32, 64, 128, 256, 512, 1,024, and 2,048, always scoring only the complementary chains. Thus 512 remains a conservative point on a sensitivity curve, rather than the method's definition.

Table 18: Exploratory ProcessBench replay: three outer splits, each with 400 score-only calibration chains (100 per public subset) and a disjoint 3,000-chain test complement. Entries are outer-split means. "Always move" uses the mixture intersection. The trust-region gate moves only when $|\operatorname{logit}(\tau)| \leq 1$, or when the fitted density has two modes and a valley at most half the height of the weaker mode. The gate itself was explored after ProcessBench was visible.

PRM	Fixed	Always move	Δ	Gate Δ	Gate action
Qwen-Math	73.38	68.44	-4.95	0.00	keep 0.5
Qwen-PRM800K	56.82	61.19	+4.37	+3.26	move 2/3
Math-Shepherd	28.96	39.52	+10.56	+10.56	move 3/3
RLHFlow-Mistral	28.40	32.06	+3.65	+3.65	move 3/3
RLHFlow-DeepSeek	27.14	35.94	+8.81	0.00	keep 0.5
Universal-PRM	0.00	70.41	+70.41	+70.41	move 3/3
ReasonFlux	44.83	66.75	+21.92	0.00	keep 0.5

What the replay does and does not show. The estimator is already practically stable for several models at 32–128 score-only chains; Math-Shepherd is less stable and becomes broadly positive at 128–256 chains. Thus 512 is a conservative diagnostic batch, not an intrinsic requirement. The unconditional result also makes abstention necessary: Qwen-Math loses 4.95 points on average across the three outer splits. A density-only gate that requires an exceptionally deep two-mode valley blocks this loss but also declines five improvements. The trust-region gate in table 18 blocks the loss and retains four improvements, while intentionally declining the two largest remaining corrections. This trade-off cannot be inferred from score geometry alone as a theorem; target logits admit many possible label assignments.

The required confirmatory protocol is therefore to predeclare the score-only batch size, chain-minimum feature, mixture specification, gate, models, metric, and bootstrap units; score a genuinely unseen target; fit only on its unlabelled calibration split; and evaluate once on a disjoint labelled test split. Until that protocol is run, this section is evidence that a useful label-free candidate boundary may exist, not evidence of a universally reliable target calibration method.

I ARTIFACTS

The anonymized supplementary material contains the code and frozen artifacts behind every table and figure: the audited public-PRM decision curves and campaign manifests; the public-PRM ProcessBench and PRMBench score caches (identifiers, labels, and scores only); the fixed-support intervention records with the 9B paired bootstrap; the Llama-3.1-8B pilot; the mechanism, few-shot, source-transfer, and fixed-pool checks; the score-only target-adaptation replay; and the prevalence-quantile analyses, including the rebuilt-backbone re-run. Its README maps each table and figure to the script that builds it and the frozen output that script reads. Campaign manifests pin benchmark-file hashes, adapter source hashes, model identifiers, and context policies, and each audited cell records the hash of its source scores. No Best threshold is fed back into a checkpoint, adapter, or

score file, and the test-label Best thresholds are kept separate from the source-transfer thresholds and from the exploratory score-only replay. Model checkpoints, third-party benchmark data, and per-example MedPRMBench score caches are not included; the fixed-support checkpoints and raw caches were not retained.

J Extended Related Work

Process supervision and step-level judgment. Outcome verifiers rerank sampled solutions (Cobbe et al., 2021); process supervision instead scores every step (Uesato et al., 2022; Lightman et al., 2024), and Monte-Carlo rollouts make step labels cheap (Wang et al., 2024). Step judgments also drive self-correction and test-time search: locating the first error lets a system backtrack to the last correct step (Samanta et al., 2026; Chitale et al., 2026). ProcessBench, PRMBench, MedPRM-Bench, and AgentProcessBench evaluate step judgments in mathematics, medicine, and tool use (Zheng et al., 2025; Song et al., 2025; Wu et al., 2026; Fan et al., 2026); supervision beyond the first error has been proposed for reflective long chains (Yang et al., 2025).

Improving the scorer. Most work on weak step judgment changes the scorer. Label noise from Monte-Carlo estimation is addressed by consensus filtering (Zhang et al., 2025b), self-denoising annotation (Ding et al., 2025), and noise-aware training (Xie et al., 2026); other work trains without step labels (Sun et al., 2025; Gadetsky et al., 2026) or learns step credit from outcomes (Yuan et al., 2024; Jia et al., 2026). Supervision targets include Q-value rankings (Li & Li, 2025), trajectory-aware rewards (Zou et al., 2025), error-typed hierarchical labels (Pala et al., 2025), tree-guided fidelity-aware labels (Zhang et al., 2025a), combined correctness and potential rewards (Wu et al., 2025), learned reliability (Li et al., 2026), and retrieval augmentation (Zhu et al., 2025). Readers generate reasoning before a verdict (Zhao et al., 2026; Khalifa et al., 2026; She et al., 2025; Lee et al., 2026), score in both directions (Zhang et al., 2026), or work in latent space (Yu et al., 2025; Ni et al., 2026; Guo et al., 2026; Liang et al., 2025; Chen et al., 2026; Alvarez & Baheri, 2026). PRISM links PRM errors to imbalanced step supervision and repairs the score function by contrastive retraining (Agrawal et al., 2026); we instead separate score quality from operating point on frozen scores and matched runs. Every such scorer still needs an explicit decision rule, and we analyze that rule rather than propose a PRM.

Calibration, thresholds, and class priors. Probability calibration is well studied, including for PRMs (Guo et al., 2017; Park et al., 2025; Ma et al., 2026); we do not measure posterior reliability, and temperature scaling leaves the zero-margin decision unchanged. Calibrated scores need not have an F1-optimal threshold of 0.5 (Lipton et al., 2014). Class-dependent logit shifts and imbalance-aware losses are classical (Menon et al., 2021; Lin et al., 2017; Cui et al., 2019), and long-tailed recognition showed that good representations can be limited by their classifier (Kang et al., 2020); our controlled runs show a PRM analogue in which the training prior moves the boundary. Prior correction under label shift is classical (Saerens et al., 2002; Lipton et al., 2018); Q_π is its rank-space form. For generative verifiers, strictness, the tendency to be under- or over-critical, is encoded in hidden states and can be steered without fine-tuning (Zhou et al., 2026); we trace the analogous offset of scalar PRMs to the training prior and separate it from discrimination. Concurrent work finds base-rate matching near-oracle for one *judge* configuration that filters PRM training data (sVanZeno, 2026); we study the evaluation-time readout of PRMs.