

Scoring-Location Sensitivity in Process Reward Models

Jai Kumar Sharma
Virginia Tech
jaisharma@vt.edu

Abstract

Process reward models (PRMs) assign scores to reasoning steps, but an evaluator must also choose where to score. We ask whether that choice changes the predicted earliest error while the reasoning text stays unchanged. On 800 selected ProcessBench questions we compare a punctuation policy and a near-midpoint policy that insert equal numbers of within-step scoring locations, preserve every reasoning byte, and map decisions back to the original steps. At threshold 0.5 the two policies disagree on the predicted earliest error for 10.4%, 14.5%, and 14.4% of questions for Qwen2.5-Math-PRM, Math-Shepherd, and Skywork, and for the marker interfaces the disagreement exceeds a separator-only comparison on the same questions. The direct comparison was identified post hoc; a disjoint, prospectively frozen 400-question panel again shows 10.0–14.3% disagreement. Two mechanisms are at work: marker-based policies can change scores at the original step endpoints, whereas Skywork keeps every token score fixed and changes the predicted earliest error only through which positions are inspected. These are controlled sensitivity measurements, not prevalence estimates for naturally occurring segmenters. PRM evaluations should report marker rendering, separators, readout positions, the mapping to original steps, and the decision threshold.

1 Introduction

An earliest-error prediction is not determined by a PRM checkpoint and reasoning trace alone. The evaluator must also decide where to insert scoring markers or which token positions to inspect; we call this rule the *scoring-location policy*. Would the same model flag the same original step if only that policy changed while the reasoning text stayed fixed?

The question matters when PRM evaluations are compared. ProcessBench responses were flattened

and resegmented before annotation (Zheng et al., 2025), marker conventions differ across scoring interfaces (Qwen Team, 2025; Zhang et al., 2025; Wang et al., 2024), and a recent survey lists standardized step definitions and evaluation protocols among the open problems (Zheng et al., 2026). We therefore compare two controlled policies on the same responses: one adds scoring locations at punctuation boundaries, the other places the same number of locations near the token midpoints of the same eligible steps. Both preserve the reasoning bytes, and every predicted error is mapped back to the original ProcessBench step. Figure 1 shows a response on which the two policies flag different original steps.

On the original 800-question panel the policies disagree on 10.4%, 14.5%, and 14.4% of questions for Qwen2.5-Math-PRM, Math-Shepherd, and Skywork, and for the marker interfaces the disagreement exceeds a separator-only comparison on the same questions. This direct comparison was identified post hoc on the original panel; we then froze it before scoring a disjoint 400-question panel, which also showed disagreement. The preregistered primary asked a different question, whether one policy differs from the reference more often than the other, and was null; §5.2 explains why that result coexists with direct policy disagreement.

We also ask how the decisions change. With marker-based interfaces, changing scoring locations can alter scores at the original endpoints. With Skywork, the token scores stay fixed, yet inspecting different positions still changes the predicted earliest error. The study measures sensitivity under controlled policies on selected questions, not its prevalence under naturally occurring segmenters; its practical implication is that a PRM evaluation should specify the scoring-location policy as well as the checkpoint.

Our contributions are (1) a controlled audit that preserves the reasoning text, matches the number of

released step 1 (annotated error), with the two cuts marked:

Adding all the kittens together gives us the total number of kittens the Doubtfire family now «S-U» has. «S-P» The total number of kittens is 21 + 12 = 33.

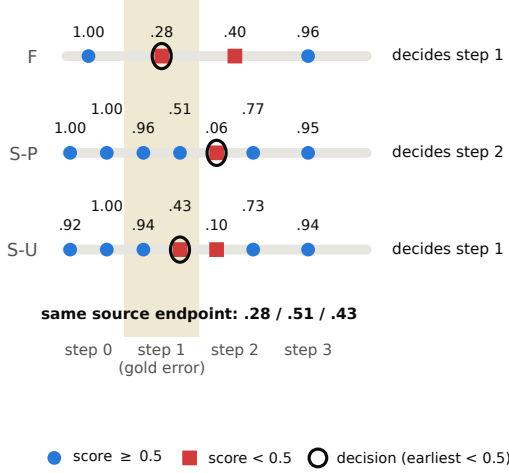


Figure 1: The same reasoning response yields different earliest-error predictions (ProcessBench row gsm8k-118, Qwen2.5-Math-PRM, fp32 canonical rescore). Top: released step 1 (the annotated error) with the punctuation cut (S-P) and the near-midpoint cut (S-U) marked at their recorded offsets. Bottom: block scores under F, S-P, and S-U; circles are scores at or above $\tau=0.5$, squares below, and the ring marks the earliest below-threshold block (the decision). At the highlighted original step-1 endpoint Qwen assigns 0.28 under F, 0.51 under S-P, and 0.43 under S-U, so F and S-U flag step 1 while S-P flags step 2: the original endpoint score changed although the reasoning text did not.

scoring locations, and maps predictions back to the original steps; (2) measurements of direct policy disagreement with separator and local-position controls and a prospectively frozen disjoint panel; and (3) a diagnostic that separates changes to existing scores from changes caused by reading different positions of a fixed score sequence, with a reporting recommendation.

2 Related Work

ProcessBench, PRMBench, and Hard2Verify establish step-level verification benchmarks (Zheng et al., 2025; Song et al., 2025; Pandit et al., 2026), building on PRM800K’s process supervision (Lightman et al., 2024).

Closest to us, EST-PRM stress-tests frozen PRMs with label-preserving structural transformations (Shihab et al., 2026), BiPRM perturbs ProcessBench step structure and reports F1 degradation in the perturbed representation (Zhang et al., 2026), and Beyond the First Error studies PRMs for

reflective chains under differing step-segmentation conventions (Yang et al., 2025). Prior PRM robustness studies examine distribution shift, formatting, and adversarial perturbations, with findings that depend on the model and the perturbation type (Dontsov et al., 2026; Tiwari et al., 2026); more broadly, FormInv studies the semantic invariance of mathematical reasoning benchmarks under paraphrase (Thomas and Thomas, 2026). These studies do not hold the reasoning bytes fixed with exact source pullback, and they evaluate perturbations by aggregate reward, PRM@N, F1, semantic-consistency, or ranking metrics rather than by source-aligned paired earliest-error disagreement.

Training-time work varies the supervision unit (Liu et al., 2025; Hu et al., 2025; Cao et al., 2025; Wang et al., 2026). Token-level value models remove explicit markers but still need a declared rule for turning token scores into a discrete earliest-error decision (Chen et al., 2025; Lee et al., 2025; He et al., 2024), and uPRM interleaves trainable marker tokens (Gadetsky et al., 2026). Generative or reasoning-style verifiers (She et al., 2025; Zhao et al., 2026; Khalifa et al., 2026; Zhong et al., 2026) and policy-aware discriminative training (Agrawal et al., 2026) change how correctness signals are produced; our measurements cover only the three scalar interfaces below. Our audit differs from the closest prior work in holding the reasoning bytes fixed, comparing source-aligned earliest-error decisions directly between two count-matched scoring policies, adding a locally matched control, and diagnosing whether existing scores or only the inspected positions changed.

3 Audit Framework

3.1 Views, byte identity, provenance

A *released step* is one of the ProcessBench steps, a *scoring location* is a position at which a PRM score is obtained, and a *view* is one placement of scoring locations over a response. The reported comparisons use four views per response. The reference view, **F**, scores the released step boundaries. The separator control, **R0**, uses those same boundaries under the model-card separator convention. The **punctuation view**, S-P, adds one scoring location at an eligible clause boundary of every jointly admissible step, and the **near-midpoint view**, S-U, adds one location to the same steps at the admissible non-punctuation whitespace boundary nearest the token midpoint. F, S-P, and S-U preserve the

original reasoning bytes once the scoring markers are removed; R0 is the explicit exception, since it changes the separator rendering. All cuts are chosen once, without labels or rewards, and reused verbatim by every interface, so S-P and S-U have identical marker counts and split-step sets. For the marker interfaces the rendered model input changes when markers are added or moved; for Skywork one unchanged input is read at different positions, and F and R0 coincide (no separator intervention exists). Every visible block records its source-coordinate span and released-step interval (coordinate conventions, admissibility floors, fail-closed invariants, and the no-truncation audit: Appendix A; literal rules: Table 5).

Roles of the views. The split pair is a controlled stress test built for identification, not a sample from deployment policies; R0 supplies a realistic separator-only anchor, and naturalistic policy prevalence is outside this study (§9).

3.2 Decisions and conservative compatibility

For each view we predict the earliest visible block whose score is below 0.5; when no block is below the threshold the output is BOT, meaning “no error detected”, not that the solution is correct. The flagged block is then mapped back to its original ProcessBench step. In the main F/S-P/S-U comparisons every flagged block maps to one original step, so two policies disagree exactly when they flag different original steps or when only one detects an error; the general rule below also handles multi-step blocks. Formally, a non-BOT decision maps through provenance to its inclusive released-step interval. Two decisions are *definitely incompatible* exactly when BOT status differs, or both are non-BOT and the intervals are disjoint; overlap counts as compatible. Formally, $I_x(V, W)$ indicates that event for row x , and the fixed-panel turnover is $\hat{d}(V, W) = n^{-1} \sum_i I_{x_i}(V, W)$; we reserve Pr for explicitly declared resampling targets.

3.3 Three scoring interfaces

Two interface families realize a scoring location differently. For Qwen and Math-Shepherd a scoring location requires a marker in the model input, so adding or moving locations changes the rendered input; for Skywork a scoring location selects a position in one already-computed token-score sequence, so the input never changes. **Marker token (Qwen).** Qwen2.5-Math-PRM-7B (Qwen Team,

2025) inserts its dedicated `<extra_0>` marker token after each block and reads the positive-class probability there. **Tag string (Shepherd).** Math-Shepherd (Wang et al., 2024) appends its tag string (space-prefixed, per the model card) and reads $P(+)/(P(+) + P(-))$ at each tag. In both, adding a scoring location changes the model input and later token history: the intervention is *adding scoring locations under a marker-input interface*, not count in isolation. **Readout-only (Skywork).** Skywork-o1-Open-PRM-1.5B (He et al., 2024) exposes a per-token value head, so all views share one byte-identical input and one forward pass, and a policy is an ordered readout set over that fixed field; readouts are mapped at token granularity, so this arm measures token-grid sensitivity. We use Skywork as a fixed-field mechanism control, not as a competitive verifier (rendering conventions and readout mapping: Appendix A; literal separators, tag strings, prompt serialization, checkpoint revisions, and cut rules: Table 5).

3.4 The fixed-field invariant

Suppose every original scoring location is retained and its score is unchanged. Inspecting additional locations may reveal an earlier low score, but it cannot remove a low score that was already present; so if the earliest detected error moves later or disappears, the score at the old decision endpoint must have changed. This directional statement needs the original endpoint set to be contained in the new one; it does not order S-P against S-U, whose added positions are not nested. Formally, with a fixed score field, adding readout positions can only preserve an existing earliest crossing or move it earlier; a move to a *later* endpoint or to BOT therefore certifies that a score at an existing endpoint changed (Lemma 1 and the upward-crossing certificate, Corollary 1, with proofs: Appendix A). Figure 1 is such a certificate case. This invariant is what lets the audit separate input-side field change from added-readout effects.

4 Study Design and Epistemic Status

Two panels. From the four ProcessBench subsets (GSM8K, MATH, OlympiadBench, Omni-MATH) we select 800 unique question keys, 200 per subset, that pass frozen eligibility gates (4–12 nonempty steps, bounded rendering length, joint split admissibility), decontaminated from a 20-response discovery pilot; selection is by pre-committed salted

Interface	S-P vs. S-U		F vs. R0
	Original (800)	Disjoint (400)	Original (800)
Qwen	83 (10.4%)	44 (11.0%)	40 (5.0%)
Shepherd	116 (14.5%)	57 (14.3%)	51 (6.4%)
Skywork	115 (14.4%)	40 (10.0%)	n/a

Table 1: Direct earliest-error disagreement between the punctuation and near-midpoint policies (S-P vs. S-U) at $\tau=0.5$ (fp32) on the original and disjoint panels, and the separator-only comparison F vs. R0 on the original panel; headers give denominators. Qwen and Shepherd are marker interfaces; Skywork is readout-only and realizes F and R0 as one readout set on one input, so no separator intervention exists (n/a).

hash, one row per key (funnel and selection diagnostics: Appendix D). A disjoint panel of 400 keys (100 per subset) was frozen before any of its rows were scored. Macro F1 of the reference view lands within about four F1 points of published figures for both marker interfaces, a coarse implementation sanity check (Appendix B).

Analysis history. The central descriptive quantity is direct policy disagreement, $\hat{d}(S-P, S-U)$, compared with the separator-only rate $\hat{d}(F, R0)$ on the same questions. The preregistered primary was a different quantity, the difference of marginal reference-relative displacement rates $\hat{d}(S-P, F) - \hat{d}(S-U, F)$. The direct comparison was identified post hoc on the original panel and then frozen before the disjoint panel was scored; “frozen” means specified before that scoring run, not before any original-panel result was known. Exact finite-panel counts come first, intervals are descriptive sensitivity analyses, and Appendix D summarizes the amendment, deviation, and status record.

Numerical precision. All headline descriptive rates use a canonical rescore in 32-bit floating point (fp32); the preregistered primary stays on its locked bfloat16 (bf16) basis, and bf16 is otherwise retained as a sensitivity (Appendix C).

5 Scoring Policies Disagree on 10–15% of Earliest-Error Decisions

5.1 Direct policy disagreement

Original panel. Under the conservative compatibility rule, the two equal-count policies produce incompatible earliest-error decisions at the preregistered threshold 0.5 on **83/800 (10.4%)** questions for Qwen, **116/800 (14.5%)** for Shepherd, and **115/800 (14.4%)** for Skywork (Table 1). The

separator-only change alone alters decisions on 40/800 (5.0%) and 51/800 (6.4%) questions for the marker interfaces, and direct S-P/S-U disagreement exceeds that comparison on the same questions by +5.4 and +8.1 percentage points (paired bootstrap intervals $[+2.9, +7.9]$ and $[+5.6, +10.6]$): a paired comparative rate difference, not an additive causal decomposition. For the readout-only interface no separator comparison exists, and its 14.4% is attributable by construction to readout selection (§6).

Disjoint panel. The disjoint panel (100 per subset, frozen before scoring, 400/400 scored) reproduces the effect: 44/400 (11.0%), 57/400 (14.3%), and 40/400 (10.0%), with the control patterns recurring; Skywork’s point estimate is numerically lower. As an internal prospective replication (same benchmark, checkpoints, code, team) it is not independent confirmation, and the effect sizes are not claimed to be identical.

Robustness and its limits. Two qualifications apply. Threshold sensitivity differs across interfaces: 70 of 83 Qwen and 90 of 115 Skywork disagreements persist at every threshold in $[0.48, 0.52]$, against only 36 of 116 for Shepherd, whose disagreement mass is threshold-proximal. And disagreement remains when cut displacement is restricted to a local token neighborhood: a matched control that relocates each punctuation cut within a ± 2 -token caliper on the canonical tokenizer, with unmatched cuts removed from both arms, gives 49/799 (6.1%), 104/799 (13.0%), and 97/799 (12.1%), so disagreement is not limited to widely separated scoring locations (Table 2); the observed incompatible pairs themselves falsify exact decision invariance on this panel. All headline descriptive rates are fp32-canonical; the largest bf16→fp32 shift is Shepherd’s direct rate (13.0%→14.5%), and the individual certified Shepherd cases differ substantially between bf16 and fp32 while its aggregate mechanism channel is stable (construction, native-grid displacement, and precision detail: Appendix C).

5.2 Why the preregistered marginal contrast missed this

Two policies can each differ from a reference equally often without agreeing with each other. As an illustration, suppose each policy changes 10 of 100 reference decisions but on disjoint sets of questions: their reference-relative rates are equal, yet they disagree with each other on 20 questions. The

	Qwen	Shepherd	Skywork
Direct, original (800)	83	116	115
Wilson 95%	[8.4, 12.7]	[12.2, 17.1]	[12.1, 17.0]
Direct, disjoint (400)	44	57	40
Wilson 95%	[8.3, 14.4]	[11.2, 18.0]	[7.4, 13.3]
Persist ± 0.02 (of orig.)	70	36	90
at $\pm 0.01 / 0.05$	77 / 49	67 / 11	101 / 46
Matched, original (799)	49 (6.1%)	104 (13.0%)	97 (12.1%)
Matched, disjoint (400)	26 (6.5%)	42 (10.5%)	37 (9.3%)
Sep.-only, disjoint (400)	16	32	n/a

Table 2: Robustness ladder at $\tau=0.5$ (fp32), grouped as direct disagreement, threshold persistence, local-position controls, and the disjoint separator-only comparison; parenthesized labels give denominators. Wilson rows are simple binomial descriptive 95% intervals in percent. “Persist” rows count original-panel disagreements that persist at every threshold in the closed band (evaluated exactly over band endpoints and score breakpoints), not new whole-panel rates. bf16 sensitivity for the disjoint separator-only cells: 21/400 for Qwen and 40/400 for Shepherd.

preregistered primary tested the difference between the two reference-relative rates, not direct disagreement, so its null result does not establish that the policies make the same decisions. Concretely, the primary compared each policy’s marginal displacement from F (exact McNemar, $\alpha=0.05$, stratified 10,000-resample cluster-bootstrap intervals, gatekept Qwen $\rightarrow$ Shepherd) and was null: RD = -0.0150 , CI $[-0.0362, +0.0063]$, $p = 0.19$ on the locked bf16 basis; under the gate, Shepherd is descriptive ($+0.0013$) and Skywork exploratory (-0.0288). The interval rules out a 10-point marginal difference but not smaller effects. The gap between the two estimands is structural. F and split-view decisions are singleton released steps or BOT, so incompatibility is the discrete metric on outcomes:

Proposition 1 (triangle bound) *For views with singleton-or-BOT decisions, write $m_V = \hat{d}(V, F)$ and $d = \hat{d}(S-P, S-U)$. Then $|m_{S-P} - m_{S-U}| \leq d \leq \min\{1, m_{S-P} + m_{S-U}\}$.*

(Proof: Appendix A.) A near-zero marginal *difference* bounds direct disagreement only from below: two policies can displace equally often relative to F while displacing different questions, and on the fp32 rescore a 16-case marginal difference coexists with 116 direct disagreements for Shepherd. The lower bound is sharp, and the upper bound is sharp whenever $m_{S-P} + m_{S-U} \leq 1$.

6 Two Mechanisms of Policy Sensitivity

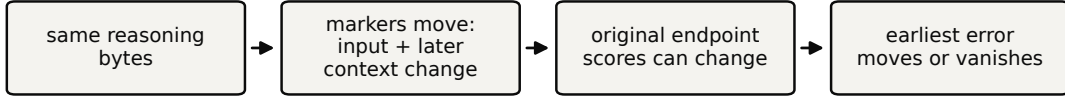
The previous section compared the punctuation and near-midpoint policies directly. This section changes the comparison: it compares each split input with the reference F while reading scores only at the original step endpoints, asking whether existing scores change enough to alter a decision; its counts do not partition the direct disagreement counts of Table 1. The interfaces expose two routes to sensitivity: marker-input policies change scores at the original endpoints (the score field moves), and the fixed-field interface changes decisions through the set of positions inspected (the readout set moves; Figure 2). The two marker-input views share marker type, count, treated steps, separators, and reasoning bytes, and the readout-only interface changes no input bytes at all, so the contrast is not explained by one arm receiving more marker perturbation.

Changing markers can change existing scores.

Recomputing each split view’s decisions at *only* the original endpoints isolates the input-side field change relative to F (it does not numerically decompose the direct S-P/S-U disagreements); Qwen shows 161 such common-endpoint decision changes under S-P (with 90 certified upward crossings), Shepherd 79 (12 certificates), and Skywork exactly 0, as the invariant requires (Table 3). A certificate is stronger than a common-endpoint change: when the earliest detected error moves later or disappears under an endpoint set that still contains the old decision endpoint, the old sub-threshold score must have risen to at least the threshold (Corollary 1); not every common-endpoint change is a certificate. Measured continuously over all shared endpoints, Qwen’s score changes are small in the median but much larger in the upper decile (median 0.006, p90 0.21); Math-Shepherd’s median absolute changes are small (0.013–0.015) while its disagreements are more threshold-sensitive (§5.1); Skywork’s changes are exactly zero, verified on the stored rewards. The certificates prove the field changed at pre-existing endpoints without separating marker history from separators, retokenization, or position, so we call the mechanism input-side scoring-interface interference.

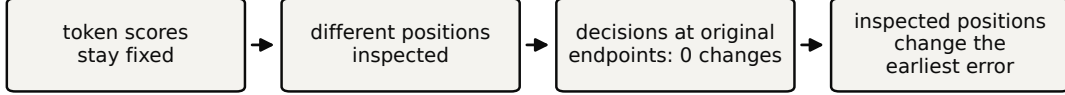
Changing readout positions can change decisions with fixed scores. Skywork is a mechanism control, not a competitive verifier at $\tau=0.5$: the threshold approximately bisects its released-

A Marker-input channel (Qwen, Math-Shepherd)



decision changes at original endpoints vs. F under S-P: Qwen 161, Shepherd 79 of which certified upward crossings: 90 and 12

B Readout-only channel (Skywork)



all disagreement comes from which positions of a fixed score sequence are inspected

Figure 2: The two channels, each compared with the reference F. Marker-input policies (top) change the rendered input, so scores at the original endpoints can change: 161 (Qwen) and 79 (Shepherd) questions change their decision when the split-input rewards are read at only the original endpoints, of which 90 and 12 are certified upward crossings (Table 3). The readout-only policy (bottom) keeps one bitwise-fixed token-score sequence, so decisions at the original endpoints never change (exactly 0) and all disagreement comes from which positions are inspected.

Interface	Changed vs. F		Certified	
	S-P	S-U	S-P	S-U
Qwen	161	139	90	80
Shepherd	79	72	12	43
Skywork	0	0	0	0

Table 3: Reference-relative mechanism diagnostics (fp32, $\tau=0.5$, original $n=800$ panel): numbers of questions whose decision, recomputed from the split-input rewards at only the original endpoints, differs from F (“Changed vs. F”), and the subset that are certified upward crossings. These are separate diagnostics, not additive parts of Table 1; Skywork’s zeros are structural under the fixed-field realization.

step-end values (55.5% below τ against 20.3% and 17.9% for the marker interfaces). Its token scores are bitwise fixed, yet changing the set of inspected positions changes the earliest-error decision on 14.4% of questions: field invariance does not imply readout-set invariance, a property of the evaluated readout policy rather than a defect of token-level value modeling. The punctuation and near-midpoint policies do read different portions of the field (63.4% vs. 55.3% below threshold), but the mechanism claim requires only that the field is fixed and the actions differ. Direct disagreement remains 11.4% under a non-overshooting floor mapping, disfavoring a pure token-overshoot artifact (Appendix C).

7 Evaluation Consequence and Scope

Changing the scoring policy also changes which response is selected from a shared candidate pool. In *up-to-eight* common-candidate pools (identical candidates within each interface, ranked per policy by minimum block reward with ties broken by the lowest candidate index; 167 questions), the selected response differs on 104/167 (62.3%), 56/167 (33.5%), and 88/167 (52.7%) questions for Qwen, Shepherd, and Skywork; restricting to pools that retain all eight candidates gives 86/132, 45/132, and 70/131 (Table 4). This is selected-response disagreement, not evidence of an answer-accuracy change.

On the analyzed answer-defined subsets ($n=64/62/64$) we observe no question on which the selected response is correct under only one policy, so on those subsets the analysis establishes selection instability rather than an answer-quality change; pool composition, the margin ladder, the similarity heuristic, and Skywork’s documented arithmetic-mean aggregation are in Appendix C. Across the reference and split policies, pooled ProcessBench-style harmonic F1 (0–1 scale) also varies by up to 0.046 with cohort- and interface-dependent sign, and clean-prefix retention moves with the policy (Appendix B).

8 Conclusion

An earliest-error prediction depends on the scoring-location policy as well as the PRM checkpoint.

	Qwen	Shepherd	Skywork
Selected response differs (167)	104	56	88
Full-eight pools only	86/132	45/132	70/131

Table 4: Selected-response disagreement under up-to-eight common-candidate selection (fp32, S-P vs. S-U, minimum aggregation, lowest-index tie rule). Margin, similarity, and answer analyses with their conditioning sets: Appendix C.

On the selected original panel, two count-matched policies over unchanged reasoning text disagree on roughly 10–15% of questions, and a prospectively frozen disjoint panel also shows disagreement. Marker-based policies can change scores at existing endpoints, while a readout-only policy can change decisions by inspecting different positions of a fixed score sequence. These findings support a concrete reporting requirement: evaluations should declare marker rendering, step separators, readout positions, the mapping to original steps, and the decision threshold, and compare models under matched policies; if several policies are relevant, their distribution and aggregation rule should be prespecified (Appendix D). The reported rates describe controlled sensitivity, not prevalence under naturally occurring scoring policies.

9 Limitations

These policies are controlled stress tests, not samples from deployment policies, so the rates are not prevalence estimates. The study covers a selected, jointly splittable ProcessBench cohort and three scalar interfaces; it does not establish generality to longer traces, other benchmarks, or current generative and reasoning-style verifiers (She et al., 2025; Zhao et al., 2026; Khalifa et al., 2026; Zhong et al., 2026), and score-scale reliability is a further active concern for scalar PRMs (Park et al., 2025). Reweighting to eligible-population subset frequencies (question keys or response rows) changes the headline rates by at most 1.6 percentage points (Appendix C; population scope: Appendix D).

The original direct comparison was post hoc; the prospectively frozen disjoint panel reproduced nonzero disagreement but shares the benchmark, checkpoints, code, and team. Threshold 0.5 is not calibrated across interfaces, Math-Shepherd’s disagreements are threshold-proximal, and its individual certified cases differ substantially between bf16 and fp32 while its aggregate channel is stable (Appendix C).

The preregistration, amendments, deviation log, code, and raw rewards are not publicly released, so the chronology in Appendix D rests on the authors’ internal records and the reproducibility record is the declared configuration in Table 5. The marked interventions are composite (readout set, input tokens, later context), and the sequential mechanism diagnostic is order-specific rather than a unique causal allocation. The selection analysis establishes selected-response instability, not answer-quality change, and the matched-coverage analysis is a failure to detect a residual difference, not a coverage attribution.

Ethics Statement

No private data, human subjects, or new annotation were used; the study evaluates model interfaces and is not human-performance or deployment-safety evidence.

References

- Aakriti Agrawal, Souradip Chakraborty, Armin Saghaian, Nihal Sharma, Rizal Fathony, Nam H. Nguyen, C. Bayan Bruss, Amrit Singh Bedi, and Furong Huang. 2026. The hidden bias of process reward models: PRISM for rewarding the right reasoning. *arXiv preprint arXiv:2606.09078*.
- Lang Cao, Renhong Chen, Yingtian Zou, Chao Peng, Huacong Xu, Yuxian Wang, Wu Ning, Qian Chen, Mofan Peng, Zijie Chen, Peishuo Su, and Yitong Li. 2025. More bang for the buck: Process reward modeling with entropy-driven uncertainty. *arXiv preprint arXiv:2503.22233*.
- Hongzhan Chen, Tao Yang, Shiping Gao, Ruijun Chen, Xiaojun Quan, Hongtao Tian, and Ting Yao. 2025. [Discriminative policy optimization for token-level reward models](#). In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 9546–9565. PMLR.
- Alexey Dontsov, Anton Korznikov, Andrey V. Galichin, and Elena Tutubalina. 2026. [Out of distribution, out of luck: Process rewards misguide reasoning models](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 421–435, Rabat, Morocco. Association for Computational Linguistics.
- Artyom Gadetsky, Maxim Kodryan, Siba Smarak Panigrahi, Hang Guo, and Maria Brbic. 2026. Unsupervised process reward models. *arXiv preprint arXiv:2605.10158*.
- Jujie He, Tianwen Wei, Rui Yan, Jiakai Liu, Chaojie Wang, Yimeng Gan, Shiwen Tu, Chris Yuhao Liu,

- Liang Zeng, Xiaokun Wang, Boyang Wang, Yongcong Li, Fuxiang Zhang, Jiacheng Xu, Bo An, Yang Liu, and Yahui Zhou. 2024. *Skywork-o1 Open Series*. Version 1.0.0. Model card: <https://huggingface.co/Skywork/Skywork-o1-Open-PRM-Qwen-2.5-1.5B>; checkpoint revision 98d69606595eedbdbbf0a7d28efdcd462ba6a67, accessed 2026-07-30.
- Yulan Hu, Sheng Ouyang, Jinman Zhao, and Yong Liu. 2025. Coarse-to-fine process reward modeling for mathematical reasoning. *arXiv preprint arXiv:2501.13622*.
- Muhammad Khalifa, Rishabh Agarwal, Lajanugen Logeswaran, Jaekyeom Kim, Hao Peng, Moontae Lee, Honglak Lee, and Lu Wang. 2026. *Process reward models that think*. *Transactions on Machine Learning Research*.
- Peng Kuang, Yanli Wang, Xiaoyu Han, Yaowenqi Liu, Kaidi Xu, and Haohan Wang. 2026. *Optimal aggregation of LLM and PRM signals for efficient test-time scaling*. In *The Fourteenth International Conference on Learning Representations (ICLR)*.
- Jung Hyun Lee, June Yong Yang, Byeongho Heo, Dongyoon Han, Kyungsu Kim, Eunho Yang, and Kang Min Yoo. 2025. Token-supervised value models for enhancing mathematical problem-solving capabilities of large language models. In *The Thirteenth International Conference on Learning Representations (ICLR)*.
- Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let’s verify step by step. In *The Twelfth International Conference on Learning Representations (ICLR)*.
- Yuliang Liu, Junjie Lu, Chaofeng Qu, Zhaoling Chen, Zefan Cai, Jason Klein Liu, Chonghan Liu, Yunhui Xia, Li Zhao, Jiang Bian, Chuheng Zhang, Wei Shen, and Zhouhan Lin. 2025. *AdaptiveStep: Automatically dividing reasoning step through model confidence*. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 39016–39031. PMLR.
- Shrey Pandit, Austin Xu, Xuan-Phi Nguyen, Yifei Ming, Caiming Xiong, and Shafiq Joty. 2026. *Hard2Verify: A step-level verification benchmark for open-ended frontier math*. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22502–22517, San Diego, California, United States. Association for Computational Linguistics.
- Young-Jin Park, Kristjan Greenewald, Kaveh Alim, Hao Wang, and Navid Azizan. 2025. *Know what you don’t know: Uncertainty calibration of process reward models*. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*. ArXiv:2506.09338.
- Qwen Team. 2024. Qwen2.5-Math-7B-Instruct. <https://huggingface.co/Qwen/Qwen2.5-Math-7B-Instruct>. Hugging Face model card; revision ef9926d75ab1d54532f6a30dd5e760355eb9aa4d, accessed 2026-07-30.
- Qwen Team. 2025. Qwen2.5-Math-PRM-7B. <https://huggingface.co/Qwen/Qwen2.5-Math-PRM-7B>. Hugging Face model card; revision 0610740060112df12585d00a1c5f4624d2f59051, accessed 2026-07-30.
- Shuaijie She, Junxiao Liu, Yifeng Liu, Jiajun Chen, Xin Huang, and Shujian Huang. 2025. *R-PRM: Reasoning-driven process reward modeling*. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 13438–13451, Suzhou, China. Association for Computational Linguistics.
- Ibne Farabi Shihab, Fariya Afrin, Sanjeda Akter, and Anuj Sharma. 2026. EST-PRM: Stress-testing process reward models before they become load-bearing. *arXiv preprint arXiv:2606.00437*.
- Mingyang Song, Zhaochen Su, Xiaoye Qu, Jiawei Zhou, and Yu Cheng. 2025. *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)*, pages 25299–25346, Vienna, Austria. Association for Computational Linguistics.
- Nishal Thomas and Noel Thomas. 2026. FormInv: A measurement protocol for semantic invariance in mathematical reasoning benchmarks. *arXiv preprint arXiv:2605.29001*.
- Rishabh Tiwari, Aditya Tomar, Udbhav Bamba, Monishwaran Maheswaran, Heng Yang, Michael W. Mahoney, Kurt Keutzer, and Amir Gholami. 2026. Reward under attack: Analyzing the robustness and hackability of process reward models. *arXiv preprint arXiv:2603.06621*.
- Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2024. *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)*, pages 9426–9439, Bangkok, Thailand. Association for Computational Linguistics.
- Teng Wang, Zhangyi Jiang, Zhenqi He, Hailei Gong, Shenyang Tong, Wenhan Yang, Zeyu Li, Yanan Zheng, Zifan He, Zewen Ye, Shengjie Ma, and Jianping Zhang. 2026. *Towards hierarchical multi-step reward models for enhanced reasoning in large language models*. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 565–576, San Diego, California, United States. Association for Computational Linguistics.

- Zhaohui Yang, Chenghua He, Xiaowen Shi, Shihong Deng, Linjing Li, Qiyue Yin, and Daxin Jiang. 2025. [Beyond the first error: Process reward models for reflective mathematical reasoning](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 4711–4728, Suzhou, China. Association for Computational Linguistics.
- Lingyin Zhang, Jun Gao, Xiaoxue Ren, and Ziqiang Cao. 2026. [The bidirectional process reward model](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12564–12580, San Diego, California, United States. Association for Computational Linguistics.
- Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2025. [The lessons of developing process reward models in mathematical reasoning](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10495–10516, Vienna, Austria. Association for Computational Linguistics.
- Jian Zhao, Runze Liu, Kaiyan Zhang, Zhimu Zhou, Junqi Gao, Dong Li, Jiafei Lyu, Zhouyi Qian, Biqing Qi, Xiu Li, and Bowen Zhou. 2026. [GenPRM: Scaling test-time compute of process reward models via generative reasoning](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(41):34932–34940.
- Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2025. [ProcessBench: Identifying process errors in mathematical reasoning](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1009–1024, Vienna, Austria. Association for Computational Linguistics.
- Congmin Zheng, Jiachen Zhu, Zhuoying Ou, Yuxiang Chen, Kangning Zhang, Rong Shan, Zeyu Zheng, Mengyue Yang, Jianghao Lin, Yong Yu, and Weinan Zhang. 2026. [A comprehensive survey of process reward models: Data generation, model construction, and usage](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3591–3607, San Diego, California, United States. Association for Computational Linguistics.
- Jianyuan Zhong, Zeju Li, Zhijian Xu, Xiangyu Wen, Kezhi Li, and Qiang Xu. 2026. [Solve-Detect-Verify: Inference-time scaling with flexible generative verifier](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 47377–47401, San Diego, California, United States. Association for Computational Linguistics.

A Formal Results and Verification

Setting. Let $s_V(p) \in [0, 1]$ be the score view V assigns to source endpoint p , endpoints ordered by source position; for a field s on an endpoint set B , let $d_s(B)$ be the earliest $p \in B$ with $s(p) < \tau$, else BOT.

Lemma 1 (fixed-field readout monotonicity)

Let $B \subseteq B'$ be finite subsets of a totally ordered source-coordinate domain and let fields s on B and s' on B' agree on the common set: $s'(p) = s(p)$ for every $p \in B$. Then $d_s(B) \neq \text{BOT}$ implies $d_{s'}(B') \neq \text{BOT}$ and $d_{s'}(B') \preceq d_s(B)$ in source order; and $d_s(B) = \text{BOT}$ implies $d_{s'}(B')$ is BOT or a new crossing in $B' \setminus B$.

Corollary 1 (upward-crossing certificate) With $B \subseteq B'$ as in Lemma 1 but s' not assumed to agree with s on B : if $d_s(B) = p \neq \text{BOT}$ while $d_{s'}(B')$ is BOT or strictly later than p in source order, then agreement fails at the old decision endpoint itself, $s'(p) \geq \tau > s(p)$.

Proof of Lemma 1. If $d_s(B) = p$ then $s(p) < \tau$, and $p \in B'$ with $s'(p) = s(p) < \tau$ by agreement on B , so the earliest sub-threshold endpoint of B' under s' exists and is at or before p . If $d_s(B) = \text{BOT}$, all of B is at or above τ under s and hence under s' , so any crossing lies in $B' \setminus B$. $\square$

Proof of Corollary 1. Suppose $d_s(B) = p \neq \text{BOT}$, so $s(p) < \tau$. If $s'(p) < \tau$ held as well, then p would be a sub-threshold endpoint of B' under s' and $d_{s'}(B')$ would be at or before p , contradicting $d_{s'}(B') = \text{BOT}$ or $d_{s'}(B') \succ p$. Hence $s'(p) \geq \tau > s(p)$. $\square$

Proof of Proposition 1. For split views and F, every decision is a singleton released step or BOT, and the frozen compatibility rule reduces to outcome inequality, so $I(V, W)$ is the discrete metric on the outcome space of released-step indices together with BOT. Write $X = I(S-P, F)$, $Y = I(S-U, F)$, $Z = I(S-P, S-U)$. Pointwise, $|X - Y| \leq Z \leq X + Y$ by the triangle and reverse-triangle inequalities for the discrete metric; then $|\mathbb{E}X - \mathbb{E}Y| \leq \mathbb{E}|X - Y| \leq \mathbb{E}Z$ and $\mathbb{E}Z \leq \mathbb{E}X + \mathbb{E}Y$. The restriction to singleton outcomes is essential: interval-overlap compatibility for wider decision intervals is reflexive and symmetric but not transitive, so the bound does not extend to it. $\square$

Rendering and readout conventions. The punctuation and near-midpoint cuts require ≥ 8

canonical-tokenizer tokens per side and fall outside paired $\text{T}_{\text{E}}\text{X}$; all split geometry is defined under the frozen Qwen2.5-Math-PRM-7B tokenizer and attributed back to Unicode code-point source coordinates under fail-closed coverage checks (“byte-identical” refers to the UTF-8 encoding covered by the pinned hashes). Math-Shepherd’s tag is rendered space-prefixed per its model card; Skywork readouts use the token whose offset span contains each block’s last character, because BPE merges block-final punctuation with following separators, and its release documents newline step-token readouts with arithmetic-mean Best-of- N aggregation (He et al., 2024), so our interior readouts are off-reference stress tests of the same value field.

Verification invariants. Fail-closed checks enforce exact-once character coverage, monotone intervals, marker/block counts, pinned input hashes, identical marker-stripped text (R0 excepted, whose separator bytes differ by design), and reconstruction of every decision from raw rewards; 800 rows $\times$ 5 views (F, R0, S-P, S-U, and M0, an even-adjacent-merge control view whose analyses are not reported here) were byte-verified per interface with zero interface-specific cuts, and no input was truncated. Every row has equal S-P/S-U unique token-readout counts with zero collisions under both readout mappings. Complete revisions, hashes, and the full deviation record are retained in the authors’ internal archive, which is not publicly released; the executed conventions are declared in Table 5.

B Evaluation Consequence Details

Metrics. ProcessBench-style harmonic F1 is the harmonic mean of two class accuracies: on gold-error trajectories a prediction is correct iff its released step equals the annotated first error, and on all-correct trajectories iff it is BOT; the value is undefined when a class is empty. Macro F1 averages the per-subset values over GSM8K, MATH, OlympiadBench, and Omni-MATH; pooled F1 computes the two class accuracies over the pooled selected cohort. The two aggregate differently and need not agree. Gold-error retention is the fraction of gold-error rows whose accepted clean prefix (the released steps strictly before the flagged block; the whole response under BOT) still contains the annotated first error, so larger values are worse.

Item	Declaration, as executed
Checkpoints	Hugging Face repositories at the pinned commit revisions: Qwen/Qwen2.5-Math-PRM-7B at 0610740060112df12585d00a1c5f4624d2f59051; peiyi9979/math-shepherd-mistral-7b-prm at 45dc0a3c9ec699b645085c098ed38dc99fba4617; Skywork/Skywork-o1-Open-PRM-Qwen-2.5-1.5B at 98d69606595eedb0bbf0a7d28efdc462ba6a67; candidate generator Qwen/Qwen2.5-Math-7B-Instruct at ef9926d75ab1d54532f6a30dd5e760355eb9aa4d. Tokenizers are loaded at the same revisions.
Qwen input	Pinned chat template with system prompt “Please reason step by step, and put your final answer within <code>\boxed{}</code> .”, user turn = problem, assistant turn = rendered response (no generation prompt). <code><extra_0></code> immediately after every visible block; F glue <code>\n\n</code> after each non-final marker at a released-step edge and no glue at a within-step cut; R0 = markers only, no glue (model-card rendering). Reward = softmax positive-class probability at each marker position.
Shepherd input	Plain text problem + “ ” + rendered response. Tag = one space followed by the two Cyrillic letters U+043A U+0438 (the model card’s “ki” tag; space-prefixed, tag token id 12902) immediately after every block; F glue <code>\n\n</code> , R0 glue <code>\n</code> (model-card newline-separated steps). Reward = $P(+)/(P(+)+P(-))$ from the logits of token ids 648 (+) and 387 (−) at each tag position.
Skywork input	One string per row: <code>< im_start > + problem + \n + response + \n</code> , response blocks joined by the original <code>\n\n</code> ; no markers; the full string is tokenized once with offset mapping. Value = sigmoid of the per-token value head. Containing mapping: the token whose offset span contains each block’s last character; floor mapping: the last token whose span ends at or before it.
S-P cut	After ., ?, or ! immediately followed by whitespace; outside balanced $\text{\TeX}$ spans; ≥ 8 pinned Qwen2.5-Math-PRM-7B tokenizer tokens on each side; minimize token imbalance, lowest character offset on ties; the whitespace belongs to the right child.
S-U cut	A whitespace run outside balanced $\text{\TeX}$, neither preceded nor followed by .?!; ≥ 8 tokens per side; minimize $ \text{left} - \text{right} $ tokens (token-midpoint rule), earliest offset on ties. A step is split iff both cuts exist (joint admissibility), exactly one cut per such step, so S-P and S-U have identical split-step sets and marker counts. Matched arm: each S-P cut relocated to the nearest admissible non-punctuation whitespace within ± 2 canonical tokens.
Cohort gates	E1 valid label and 4–12 nonempty released steps; E2 R0 full-chat rendering $\leq 2,048$ pinned-tokenizer tokens; E3 at least one jointly admissible step; 26 pilot-contaminated rows removed; one row per question key; 200 keys per subset by a pre-committed salted SHA-256 key over the subset, question key, and row id. Context limit 4,096 tokens; no input truncated.
Decision rule	Earliest visible block with reward < 0.5 , else BOT; two decisions are incompatible iff BOT status differs or both are non-BOT with disjoint inclusive released-step intervals.
Numerics	fp32 canonical rescore with bf16 as a sensitivity; transformers 4.48.3, Python 3.10.20, one NVIDIA RTX A6000; the rendered-text SHA-256 of every row is asserted equal across passes.

Table 5: Interface declaration: the rendering, readout, cut, and selection conventions needed to reproduce the scored inputs, i.e., the quantities this paper asks PRM evaluations to report.

Implementation sanity check. Macro F1 of the F view is 69.9 (original panel) and 76.7 (disjoint panel) for Qwen2.5-Math-PRM against a published 73.5, and 28.8 and 29.3 for Math-Shepherd against a published 31.5 (Zheng et al., 2025; Zhang et al., 2025): within about four F1 points in every cell, a coarse check on the scoring-interface implementation only.

Panel	Interface	F	S-P	S-U
Original	Qwen	0.699	0.673	0.693
Original	Shepherd	0.281	0.298	0.261
Original	Skywork	0.336	0.322	0.295
Disjoint	Qwen	0.772	0.726	0.732
Disjoint	Shepherd	0.279	0.325	0.289
Disjoint	Skywork	0.361	0.339	0.342

Table 6: Pooled ProcessBench-style harmonic F1 (0–1 scale; fp32; $\tau=0.5$; post hoc, not leaderboard-comparable) by panel, interface, and view.

Pooled harmonic F1. Across the reference and split views (Table 6) the pooled value varies by up to 0.046 (disjoint Qwen, F vs. S-P), with cohort- and interface-dependent sign; the largest direct S-P vs. S-U difference is 0.037 (original Shepherd).

Clean-prefix retention. In a clean-prefix use (accept the released steps strictly before the flagged error), gold-error retention (larger is worse) on the 504 gold-error rows moves with the policy at $\tau=0.5$: 0.276/0.361/0.321 for Qwen under F/S-P/S-U, 0.742/0.714/0.760 for Shepherd, and 0.145/0.115/0.113 for Skywork. These are operating-point differences. At matched accepted-character coverage the S-P–S-U gold-error-retention gap is -0.008 $[-0.023, +0.009]$ (Qwen), -0.003 $[-0.015, +0.009]$ (Shepherd), and -0.002 $[-0.015, +0.010]$ (Skywork), replicating on the disjoint panel with every interval straddling zero. The comparison interpolates re-

tention piecewise-linearly in coverage between adjacent threshold operating points and evaluates both policies at the midpoint of their own $\tau=0.5$ coverages inside every bootstrap replicate; it is a failure to detect a residual difference under a post hoc randomized-threshold interpolation, not a coverage attribution.

C Core Robustness

Thresholds beyond the band. Under post hoc cross-panel calibration, one threshold per interface is chosen on the disjoint panel’s F view as the achievable all-correct false-positive rate nearest 0.10 (the rate is a step function of τ ; ties are broken by the smaller count) and applied unchanged to the original panel, where direct disagreement stays high: 75/99/114 of 800. This is a post hoc sensitivity, not a further prospective confirmation. Among the $\tau=0.5$ direct disagreements on the original panel (83/116/115), exactly one policy is strictly correct in 52/36/65 cases, where a decision is strictly correct if its released step equals the annotated first-error step or if it is BOT on an all-correct trajectory; policy choice therefore changes correctness, not just labels. The nine-point trapezoidal summary of direct disagreement over $\tau \in \{0.1, \dots, 0.9\}$ is 0.126/0.117/0.124, an unnormalized integral over a τ -range of width 0.8 (divide by 0.8 for the average rate) and a grid summary rather than the exact breakpoint integral. Reweighting the equal-subset rates to eligible-population subset frequencies moves each headline rate by at most 1.6 percentage points (eligible question keys: 10.4 $\rightarrow$ 11.1%, 14.5 $\rightarrow$ 15.7%, 14.4 $\rightarrow$ 13.5%; eligible response rows: 11.4%, 16.1%, 13.5%).

Matched near-position control. The near-position arm relocates each matched punctuation cut to the nearest admissible non-punctuation whitespace within a ± 2 left-token caliper (3,108 of 3,206 cuts on 799/800 rows), with unmatched cuts removed from *both* arms under a fail-closed row-level equal-count assertion (this corrects an earlier construction whose arms were not marker-count matched). The corrected contrast is 49/799, 6.1% [4.7, 8.0] (Qwen); 104/799, 13.0% [10.9, 15.5] (Shepherd); 97/799, 12.1% [10.1, 14.6] (Skywork). Relocation is token-local on every native grid (maximum 2 Skywork tokens; maximum 5 Mistral tokens, 99.7% within 3). The exploratory one-cut variant relocates a single hash-selected cut by one to two canonical tokens, retaining all other scoring

locations, and still changes 7/400 (1.8%), 17/400 (4.3%), and 20/400 (5.0%) of disjoint-panel decisions. Paired separator-vs-placement differences use a subset-stratified question-level paired bootstrap (10,000 resamples; discordant pairs 74/31 and 85/20).

Precision sensitivity. A later frozen amendment rescored both panels at fp32, the canonical basis; bf16 is retained as a sensitivity. Reward differences $|r_{\text{fp32}} - r_{\text{bf16}}|$ have medians 0.0007 (Qwen), 0.0067 (Shepherd), and 0.0026 (Skywork); the largest headline shift is Shepherd’s direct rate (13.0% $\rightarrow$ 14.5%), the other five direct headline rates move $\leq 1.2\text{pp}$, and whether a majority of Shepherd’s direct disagreements persisted across the threshold band changed under fp32 (36/116 survive ± 0.02 , from 57/104 at bf16). Qwen and Skywork preserve most disagreement identities across precisions; for Shepherd the individual certified cases differ substantially between bf16 and fp32 while its aggregate common-endpoint count is stable (79 vs. 78 for S-P), which is why the Shepherd mechanism claim rests on the aggregate channel.

Field, drift, and mapping detail. Over all 5,343 shared endpoints, Qwen’s S-P drift has median 0.006, p90 0.21, 18.5% above 0.10 (Spearman 0.96); Shepherd’s medians are 0.015/0.013 with at most 1.0% above 0.10 (Spearman 0.992); Skywork’s drift is exactly 0 at every common endpoint (Spearman 1.0). Among gold-error rows where F fires at $\tau=0.5$ (437 rows), the mean signed change at F’s predicted crossing is +0.175 (Qwen S-P) and +0.130 (S-U), consistent with erasure of existing F flags; Shepherd’s is near zero (-0.004 , +0.017; 277 rows); this signed statistic is threshold-conditioned and selection-prone, so the absolute drift above is the threshold-free evidence. The threshold approximately bisects Skywork’s released-step-end values (55.5% below $\tau=0.5$ vs. 20.3% and 17.9% for the marker interfaces), its punctuation cuts read 8.1 points more below-threshold field than its midpoint cuts (63.4% vs. 55.3%), and under the preregistered non-overshooting floor mapping direct disagreement is 91/800 (11.4%) against 115/800 (14.4%) under the containing mapping, with the corrected matched rate 85/799 (10.6%) against 97/799 (12.1%); both ratios exceed the pre-committed 0.5 persistence bar.

Up-to-eight selection diagnostics. Eight candidates per question were generated once with Qwen/Qwen2.5-Math-7B-Instruct (Qwen Team, 2024) (temperature 0.8, top- p 0.95, at most 1,024 new tokens, per-candidate seeds fixed before generation). The question population for this generation is different from the released-response cohort: it is the set of distinct ProcessBench GSM8K and MATH problems excluding the original and disjoint panels and the pilot rows, with no released-response gates (E1–E3 apply to released responses, not to newly generated candidates). This leaves 69 GSM8K problems, all of which were taken, and a larger MATH pool from which 100 were drawn by salted hash, for 169 questions; two MATH questions had no candidate scored under both policies of a contrast (the same two for every interface) and are excluded, leaving the 167 analyzed pools. Pools are policy-matched within each interface: a question’s common pool is the set of candidates scored under both policies of the contrast, candidates scored under only one policy are dropped from both rankings, and pools are therefore up to eight (eight candidates on 132 questions, 131 for Skywork, and as few as one) and are not candidate-identical across interfaces. Each policy selects the candidate with the highest aggregated score (minimum block reward for the headline; log-product and geometric mean as sensitivities), ties broken by the lowest candidate index; single-candidate pools cannot disagree, count as agreement in the 167-question rate, and are excluded from margin analyses. Selected-response disagreement conditional on both policies’ top-two margins strictly exceeding a rung (Table 7) is monotone non-increasing in every interface under all three aggregations; margins are within-interface quantities and are not comparable across interfaces.

Rung >	0.001	0.005	0.01	0.02
Qwen, % (n)	42.6 (61)	27.0 (37)	19.4 (31)	12.5 (24)
Shepherd, % (n)	27.3 (143)	17.8 (107)	9.1 (88)	4.9 (61)
Skywork, % (n)	52.2 (136)	50.0 (110)	44.2 (86)	37.7 (61)

Table 7: Selected-response disagreement (%) among questions whose top-two raw score margins under *both* policies strictly exceed the rung, with the retained n in parentheses (minimum aggregation, fp32; rungs in reward units).

Trace similarity is the `difflib.SequenceMatcher` ratio on whitespace-normalized strings. Each disagreement pair is

assigned to exactly one bucket by the frozen priority identical normalized text > answer unparseable > different parsed answers > near-duplicate (same parsed answer, ratio ≥ 0.9) > textually distinct, giving textually distinct / near-duplicate / different parsed answer / unparseable counts of 95/5/2/2 (Qwen), 46/5/0/5 (Shepherd), and 79/2/2/5 (Skywork) with no identical-text pairs; “textually distinct” is a declared string heuristic, not semantic distinctness. Answers are the last parseable `\boxed{ }` payload after whitespace normalization; the answer-defined subset requires both selected answers to parse and a gold answer to be available ($n=64/62/64$), and the answer-diverse subset requires at least two distinct parsed answers in the pool. On answer-diverse pools selected-answer disagreement is 2/27, 0/22, and 2/24; on the answer-defined subsets the paired correctness deltas are exactly 0 with 0 discordant pairs (exact one-sided 95% upper bounds $1 - 0.05^{1/n}$ on the discordance probability: 0.046/0.047/0.046; 0.044 under the $n=66$ unparseable-as-incorrect sensitivity); zero observed discordance bounds, and does not exclude, a nonzero population rate. Under the documented Skywork reference aggregation (arithmetic mean over step readouts (He et al., 2024)), computed by a procedure that first reproduces the canonical minimum-aggregation counts exactly, selected-response disagreement is 103/167 (61.7%), 43/167 (25.7%), and 87/167 (52.1%); Skywork’s rate under its own documented aggregation is nearly identical to the minimum-aggregation stress test (52.1% vs. 52.7%), and aggregation choice is an active evaluation variable even at the response level (Kuang et al., 2026).

D Epistemic Status and Scope

Amendments and deviations. Seven frozen amendments extended the preregistration (controls, disjoint panel, ranking analysis, floor mapping, fp32 rescore), each specified before its own scoring. Six deviations are logged with timestamps in the authors’ internal, unreleased record; the three affecting analyses reported here are: Shepherd’s tag rendering corrected to the model-card convention before any reward analysis, Skywork readouts mapped to the containing token before any reward was observed, and the first matched arms, which were not marker-count matched and are superseded by the corrected arms reported here.

Analysis	Status
Marginal S-P vs. S-U primary	Preregistered before scoring; null; locked to bf16
Direct disagreement	Post hoc on the original panel; prospectively estimated on the frozen disjoint panel
Matched, one-cut, band, floor controls	Frozen before their scoring runs
Precision rescore (fp32 canonical)	Preregistered full-cohort rescore; changed conclusions reported as changed
Downstream analyses (F1, coverage, selection)	Exploratory or post hoc; descriptive uncertainty

Table 8: Condensed epistemic status. Post hoc p-values are unadjusted unless stated; bootstrap intervals target repeated equal-subset draws from the four eligible strata, and eligible-population reweighting is reported separately (Appendix C).

Policy marginalization objects. For a declared, outcome-independent policy distribution μ and a per-example loss L , a marginalized benchmark reports $R_\mu(\theta) = \mathbb{E}_{(x,y)} \mathbb{E}_{\pi \sim \mu} L(D_{\theta,\pi}(x), y)$; for confusion-count metrics such as harmonic F1, per-policy averaging $\mathbb{E}_\mu[H(C_{\theta,\pi})]$ and pooled counting $H(\mathbb{E}_\mu[C_{\theta,\pi}])$ are different estimands, and the policy-draw mechanism (per benchmark run or per question) must be declared.

External validity. Selection funnel: of the 3,400 ProcessBench rows, 2,916 have 4–12 nonempty released steps (E1), 2,915 of those also render within the 2,048-token R0 ceiling (E2), and 2,842 also contain at least one jointly admissible step (E3), so the intervention is defined for 83.6% (2,842/3,400) of rows; removing 26 pilot-contaminated rows leaves the 2,816-row eligible response funnel (340/795/826/855 per subset), which collapses to 2,416 eligible question keys (321/763/511/821) under one row per key, from which the fixed panel draws 200 keys per subset by pre-committed salted hash. Reweighting to eligible-key or eligible-row frequencies is a sensitivity, not the sampling target. Selected and eligible-but-excluded rows differ modestly on score-blind features (all $|SMD| \leq 0.20$, largest on length), with selected rows somewhat shorter and slightly less error-heavy; the SMD audit uses a local proxy of the E3 gate (the authoritative gate needs the pinned tokenizer): the proxy admits 2,781 rows (331/779/824/847 per subset against the authoritative 340/795/826/855), 792 of the 800 selected rows pass it, and the comparison is between all 800 selected rows and the 1,989 proxy-admitted

rows that were not selected. The reported rates are not estimates of full-ProcessBench or naturalistic-segmenter prevalence.