Step Partitions Are Part of PRM Specifications

Anonymous authors
Paper under double-blind review

Abstract

Process reward models (PRMs) score reasoning steps, but completed solutions need not have a unique step partition. We formalize *partition well-definedness* and introduce PRM-PART to audit candidate selection while fixing the normalized ordered payload and moving only contiguous boundaries. Across three open PRMs, boundaries alter decisions beyond mechanical aggregation. In two separately frozen 1,000-problem cohorts, source-to-canonical parsing and frozen multi-view scoring change 13.2–14.7% and 25.4–43.9% of tolerance-robust winners for two PRMs; the multi-view result also transfers across two rollout generators. A selected Qwen audit retains 25/420 pools after text and AI semantic screens. Under a subsequent model-card direct-join control, boundary changes alter 13/25 winners, 11 robustly; the equal-step/equal-token subset has 9/19 changes, 7 robust. Fresh score-only search produces an actual wrong selection in one of 38 eligible eight-candidate Qwen pools (128 screened) in a separately frozen Level-5 stratum, with CPU confirmation. This stratum was motivated by the original sample's answer composition. The original 33/128 eligible pools show no harm; Level 5 also contains a repair, leaving Qwen accuracy unchanged. Equal-budget random search finds the same harmful case. Together, these results establish partition-dependent selection and an attainable correctness failure under fixed content. Reproducible scoring therefore requires a partition contract specifying normalization, atomization, ownership, serialization, readout, and aggregation.

1 Introduction

Process reward models assign credit at intermediate steps and are increasingly used for selection, filtering, and search (Uesato et al., 2022; Lightman et al., 2024). Their interface usually inserts structural markers at step boundaries, reads rewards at those markers, and aggregates the values. Yet a completed calculation and its explanation may be written in one paragraph or several without changing the operations, their order, or the final answer.

Why a changed winner matters. Choose a decimal three-digit integer uniformly at random. The probability that its base-9 and base-11 representations both have three digits is $152/225$. In a fixed eight-candidate pool, Qwen initially selects a correct derivation. A wrong candidate uses the invalid base-9 digit 9 and concludes $233/300$. Moving only its scoring boundaries raises its product score above every fixed rival, so the selected answer becomes wrong (Table 1). The candidate's text, ordered non-marker tokens, marker count, input length, and final marker position are preserved. Independent CPU scoring confirms the switch; random search also finds it.

This matters for a content-level API that promises $S(\tau)$ but computes $S(\tau, \mathcal{P})$ without specifying $\mathcal{P}$: control over segmentation can become control over which answer is selected. A step-conditioned PRM may legitimately score segmentations differently when they are part of its declared task. We therefore audit how the partition choices admitted by an interface affect selection, distinguishing syntactic admissibility, semantic-step qualification, and final-answer correctness.

Deployment scope and ownership. We study verification and reranking of *completed* traces. Candidate, middleware, or scorer may own the partition rule. Changing legal delimiters upstream makes the serializer insert different native markers and readouts; our score-only adversary sees rewards but no correctness labels. Online search states, where a boundary may define a new action time, are outside this audit.

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.

Table 1: One incorrect candidate takes over a fixed eight-candidate Qwen pool. Only its four-block partition changes; all seven rivals retain baseline scores. Each target view has the same 857 tokens apart from marker placement. This Level-5 example was chosen for illustration after outcomes; the pool, target, and search rules were frozen before scoring.

Candidate / view	Groups of original units	Product score
Wrong / baseline	[1–2] [3–4] [5–6] [7–9]	0.945896265
Wrong / adaptive	[1] [2] [3–7] [8–9]	0.999375501
Wrong / random	[1] [2–7] [8] [9]	0.999559244
Best correct / fixed	[1] [2] [3–4] [5–7]	0.998571481

Delimiter protocol versus semantic step. Qwen recommends double newlines between steps in Qwen2.5-Math-Instruct responses and inserts `<extra_0>` after each supplied step (Qwen Team, 2025). This fixes delimiter and readout syntax, but not a unique boundary for each completed solution. Coherent decompositions can share the convention; conversely, a contiguous block need not be a natural semantic step. The partition lattice defines our syntactic audit domain; semantic-step coherence is assessed separately. We distinguish source paragraphs, deterministic parsers, and extreme coarsenings below.

Prior work establishes PRM brittleness under textual and structural perturbations and studies adaptive or multi-scale segmentation (Tiwari et al., 2026; Shihab et al., 2026; Liu et al., 2025; Cao et al., 2025; Wang et al., 2025). We isolate the partition counterfactual: the atomic reasoning payload and its order are fixed, while only legal contiguous boundaries change. This separates partition choice from paraphrase, added verification, verbosity, and reordering. Our contribution is a measurement framework and empirical audit of this scoring interface:

1. We formulate a fixed-payload audit of completed-trace scoring, with score and decision invariance, attainable intervals, and exact ordering certificates on finite partition lattices.
2. PRM-PART combines executable traces, expert-annotated coarsenings, and natural candidate pools. Across Qwen2.5-Math-PRM (Yang et al., 2024), Math-Shepherd (Wang et al., 2024), and Skywork-PRM (Skywork AI, 2024), it separates aggregation mechanics from fresh-forward partition dependence and distinguishes fixed-rule disagreement from max-over-views stress tests.
3. A fresh, locally prespecified experiment separates wrong-pair reversals, adaptive-versus-random search, and full-pool correctness changes. We identify a CPU-confirmed harmful selection, quantify harm and repair across the declared strata, and demonstrate scorer-owned contracts that close the tested boundary-control channel.

2 Related work

Robustness and reward hacking. Reward Under Attack/PRM-BiasBench studies wording, verbosity, order, adversarial tokens, and reward-only optimization (Tiwari et al., 2026). Evaluator Stress Tests audits semantic edits (Shihab et al., 2025); EST-PRM adds steps, reorders dependencies, and studies mean aggregation (Shihab et al., 2026); Quality-Diversity PRM testing searches correctness-changing edits (Shihab & Afrin, 2026). PRM-PART holds the normalized ordered payload fixed and changes only contiguous boundaries, separating fresh-forward dependence from endpoint-subset aggregation. Appendix B compares interventions.

Granularity and supervision. AdaptiveStep uses confidence thresholds to place boundaries and compares test-time rating positions on completed responses: confidence versus random for ASPRM, and newline versus random for its baselines (Liu et al., 2025); EDU-PRM uses entropy-based segmentation for supervision (Cao et al., 2025), and HRM trains on merged consecutive-step segments (Wang et al., 2025). We share their premise that scoring position matters. Our audit jointly tests selected-candidate identity under fixed ordered payloads, exact ordering over a contiguous-partition lattice, fixed-forward versus fresh-forward dependence, and partition ownership. BiRM adds future success (Chen et al., 2025), Intra-Trajectory Consistency regularizes adjacent rewards (Zhou et al., 2025),

and Beyond the First Error and Cliff study semantic error boundaries (Yang et al., 2025; Han et al., 2026).

3 Partition equivalence and scoring contracts

Let $\tau = (u_1, \ldots, u_m)$ be a completed trace of ordered atomic units. A *formally admissible* partition $\mathcal{P} = (B_1, \ldots, B_k)$ divides that sequence into nonempty contiguous blocks without changing atom payload or order; $\Pi(\tau)$ is the set of all such partitions. This definition fixes the audit's structural equivalence class; semantic-step coherence is a separate qualification criterion. The implemented scoring pipeline is

$$S(\tau, \mathcal{P}) = A_{\mathcal{P}}(Q(\text{serialize}(\tau, \mathcal{P}))), \tag{1}$$

where serialization includes protocol markers, Q produces block/readout scores, and $A_{\mathcal{P}}$ aggregates a partition-dependent number of values. Parser, serialization, model, readout, and aggregation jointly specify the reward.

Score well-definedness. S descends to a scalar on completed reasoning content if $S(\tau, \mathcal{P}) = S(\tau, \mathcal{P}')$ for all $\mathcal{P}, \mathcal{P}' \in \Pi(\tau)$. Verdict invariance and same-problem ranking invariance apply a decision rule or pairwise comparison and are weaker. Localization is equivariant when a coarsened predicted error block continues to contain the same faulty atom. All of this is defined relative to a declared base atomization $\mathcal{A}(\tau)$: legal partitions can only coarsen existing atoms and cannot split within one, so $\Pi(\tau)$, $R(c)$, and well-definedness are conditional on that declared atomization (which the synthetic, expert-annotated, and natural instantiations set differently).

For candidate c, define $R(c) = \{S(c, \mathcal{P}) : \mathcal{P} \in \Pi(c)\}$ and $I(c) = [\min R(c), \max R(c)]$. The ordering $c \succ e$ is partition-certified when $\min R(c) > \max R(e)$. A reversal is attainable when $\max R(e) > \min R(c)$, which is necessary and sufficient when each candidate's partition may be chosen independently; under a shared whole-pool partition policy the same interval overlap is merely necessary, not sufficient to guarantee a reversal. This separates harmless score variation from decision-relevant interval overlap.

Empirical estimands. For an aggregation-specific development threshold t_A and $s_{i\mathcal{P}} = S(\tau_i, \mathcal{P})$, we report the partition score range $\text{PSR}_i = \max_{\mathcal{P}} s_{i\mathcal{P}} - \min_{\mathcal{P}} s_{i\mathcal{P}}$ and verdict-flip rate

$$\text{VFR} = n^{-1} \sum_i \mathbf{1}\left\{ \min_{\mathcal{P}} s_{i\mathcal{P}} < t_A \le \max_{\mathcal{P}} s_{i\mathcal{P}} \right\}.$$

Localization consistency measures how often the minimum-scoring block contains the annotated first-error atom. Conditional ASR@Q measures whether score-only partition search makes an incorrect candidate outrank its fixed, unique process-correct comparator, among pairs where it initially trails under the fine partition. It permits Q additional queries beyond fine, so ASR@16 uses at most 17 views. Natural-pool budget B counts all queried views including baseline; the fresh wrong-selection study also uses this total-budget convention. Eligible-pair counts and search details are given with the results and in Appendix O.

Proposition 1 (Fixed-forward aggregation)**.** *Suppose every atomic endpoint i has a partition-independent score r_i and a partition exposes any endpoint subset containing m. Last aggregation is always invariant. Mean is invariant iff every earlier $r_i = r_m$. Min is invariant iff $r_m \le r_i$ for every $i < m$. For positive scores, product is invariant iff every earlier $r_i = 1$.*

Proof. Compare the fully merged endpoint set $\{m\}$ with every $\{i, m\}$. The resulting equalities give the necessary conditions; substitution proves sufficiency for all larger endpoint subsets. □

Context and canonicalization. Under last aggregation, variation of the final readout across fresh forwards identifies dependence before aggregation when fixed-forward replay is invariant. Conversely, a deterministic scorer-owned canonicalizer makes declared-equivalent inputs identical and, with deterministic scoring, fixes their scores; utility and semantic-step quality remain separate evaluation criteria. These direct consequences of the scoring interface are formalized in Appendix C.

Table 2: Evidence populations and intervention status. The two 1,000-pool BiRM confirmations are distinct source allocations. Public-parser extensions reuse the second cohort after its outcomes were known.

Population	Size	Intervention and evidential role
Synthetic / expert	8,000 / 5,400 traces	Full or capped lattices; verdicts and conditional ranking
Exploratory BiRM	500×15	Frozen-view search and policy development
Reserved BiRM	$1,000 \times 15$	Frozen $B \leq 8$ search; score-only confirmation
Untouched BiRM	$1,000 \times 15$	Source/canonical confirmation; later all-pairs parser extension
MATH-500 transfer	420×8; 89×8	Two generators; common-mask model transfer, with prior Qwen results
Fresh MATH	$1,024 \times 16$	76 qualified pairs; 33 original and 38 separate Level-5 eight-candidate pools

Score invariance implies fixed-threshold verdict invariance, but a stable verdict need not preserve ranking: even small score changes can reverse a small margin. Attainable reversals characterize the feasible orderings; policy-specific experiments measure their observed frequency.

4 Evidence design

Table 2 separates exhaustive audits, frozen-view search, and deterministic parser comparisons. Exhaustion supports exact certificates for short traces; finite-view maxima measure reachability under that policy; parser pairs measure implementation disagreement. Each result is interpreted relative to its audited population and partition policy.

Executable and expert-annotated traces. PRM-PART-Synth has 300 development problems and 2,000 test problems, each with one correct and three controlled-error traces. Four to eight executable arithmetic or linear-equation atoms allow all 2^{m-1} partitions to be scored. The 1,200 fine-view development traces fix each reducer's threshold by balanced accuracy with deterministic tie-breaking, and fix standardization statistics. ProcessBench-P retains 3,400 expert-annotated traces (Zheng et al., 2025): up to ten atoms are exhaustively partitioned and longer traces use 128 fixed-seed coarsenings. PRMBench-P contains 2,000 category-stratified traces with 64 frozen partitions (Song et al., 2025). Coarsened error blocks inherit the annotated first-error atom. Ranking pairs require a unique process-correct comparator; absent final-answer labels remain unavailable.

Natural candidate allocations. BiRM pools come from a pinned candidate release (Chen et al., 2025), split by source identity. The exploratory cohort has 500 questions with 15 candidates each. Fine, fully merged, and up to six hash-seeded views preserve every candidate's ordered atoms; Qwen and Math-Shepherd share IDs and masks under their native 4,096-token limits. Before opening payloads or outcome metadata, a separate 1,000-group reserved cohort freezes eligibility, checkpoints, masks, seeds, and a no-replacement rule. All pools pass. Its primary $B = 8$ gate requires both models' bootstrap lower endpoints for robust winner change to exceed 20%. Robust changes require top-two margins $> 4 \times 10^{-5}$ under both policies. Separate answer/reference/correctness fields stay unopened through the score-only audit, although responses necessarily include final-answer text.

A second disjoint 1,000-pool cohort freezes source versus `text-canonical-v1` scoring before opening responses. The subsequent four-rule parser comparison uses these same candidates, with rules and analysis fixed before new scores but after outcomes were known. We report this post-outcome extension separately from the earlier confirmation (Appendices G and S).

MATH-500 transfer freezes a new common-mask, first-eight-candidate test before Math-Shepherd transfer scoring. Qwen-generated rollouts cover all 500 problems; DeepSeek generation samples 200. Whole-pool context gating retains 420/500 and 89/200, without replacement. This is a frozen model-transfer test using previously available payloads and Qwen-only results.

Table 3: Full-pool correctness at $B = 16$. Retained pools are out of 128 screened per stratum. Adaptive and random search have identical harm, repair, and correctness counts shown here; their winner identities can differ (Table 6). Level 5 is label-motivated and reported separately.

Stratum	PRM	Retained	Harm	Repair	Correct (before → after)
Original	Qwen	33/128	0	0	31 → 31
Original	Math-Shepherd	33/128	0	0	30 → 30
Level 5	Qwen	38/128	1	1	34 → 34
Level 5	Math-Shepherd	38/128	0	1	31 → 32

Fresh wrong-selection study. Before inspecting new generation text or PRM scores, a local protocol freezes 1,024 MATH test questions (Hendrycks et al., 2021) and 16 Qwen2.5-Math-7B-Instruct candidates per question. Score-blind mathematical review yields 76 sound-correct/actually-wrong pairs. Separate fixed pools retain candidates 0–7 and a hash-owned target, with all rivals held at baseline. A later Level-5 stratum is motivated by the original pools' answer composition but frozen before PRM scoring. We report the strata separately (Appendix A). A subsequent check restricted to reversal cases by one independent human reviewer, with PRM scores hidden, agreed with the AI judgments (Appendix A).

Records and inference. Audit records pin atoms, masks, markers, readouts, scores, checkpoints, parser/serializer hashes, and completion receipts. Normalized ordered payload is invariant; model token inputs can differ. Thresholds are never refit on test partitions. Unless explicitly stated otherwise, individual 95% intervals use 10,000 source-problem bootstrap draws, paired across policies. They are not simultaneous family-wise intervals. Unknown selected outcomes instead yield sharp sample bounds over binary completions, sharing each candidate's unknown label across policies. Scores are not numerically compared across models; score-only policies never use correctness labels.

5 Decision changes under declared partitions

Wrong selection under fixed count and length. In the fresh study, both PRMs score exactly four blocks of complete original units with the product reducer. Every legal view preserves the exact response, ordered non-marker tokens, total token count, and final marker position. Search sees only the target's scores. At $B = 16$ total queries, adaptive search makes the wrong candidate overtake its fixed correct comparator in 7/68 initially correct-leading Qwen pairs and 19/57 Math-Shepherd pairs; random search gives 9/68 and 14/57 (Table 5). Both endpoints must exceed the 4×10^{-5} margin. The paired comparisons do not establish a cross-model adaptive advantage ($p = .625$ and .0625).

The complete-pool endpoint requires the target to beat every fixed rival. Table 3 separates coverage, harm, and repair: neither policy causes harm in the original stratum; Level-5 Qwen has one harmful switch and one repair, while Math-Shepherd has one repair. CPU inference confirms Qwen's harmful case with both winner gaps above the margin. Reporting coverage, harm, and repair together distinguishes the demonstrated correctness failure from the aggregate accuracy of each declared stratum.

Native equal-count whole-pool audit. A score-blind text rule retains 28/420 Qwen-generator MATH-500 pools, each with eight candidates admitting two prose-based step lists formed by one adjacent-boundary shift. Three calibrated AI reviewers inspect all 224 candidate pairs with scores hidden; requiring all eight pairs to pass all three leaves 25 pools. On this selected set, changing boundaries under Qwen's model-card direct `<extra_0>` join and released product reducer changes 13/25 winners, 11 with both top-two log-product gaps above 4×10^{-5}. Nineteen pools also retain equal input-token counts for every candidate pair: 9 change winner, 7 robustly. Independent CPU FP32 checks confirm all 11 robust direct-join changes.

The direct-join analysis is a post-score serializer sensitivity on the same preselected, AI-qualified cohort. The original space-padded join preserves step and token counts throughout the 28-pool set and changes 17/28 winners (15 robust), or 15/25 (13 robust) after AI qualification. A post-inspection list-free sensitivity gives 4/11 changes (3 robust). Rates refer to the stated qualified cohorts; full inputs, AI judgments, and scores are in Appendix J and the supplement.

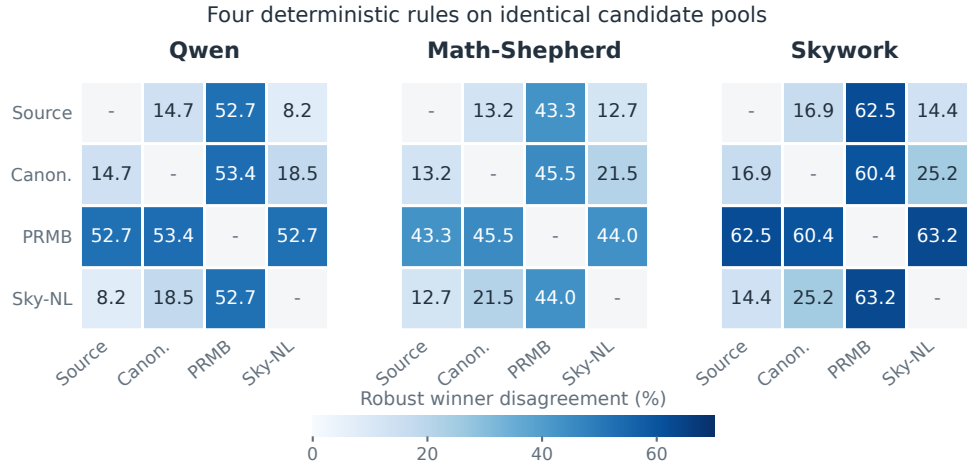

Figure 1: All six deterministic parser comparisons on the same 1,000×15 BiRM pools. Cells show robust winner disagreement (%); both top-two margins must exceed 4×10^{-5}. Each panel uses the same color scale. Source/canonical comparisons precede the post-outcome public-parser extension; all intervals and accuracy contrasts are in Table 18.

Public parser rules. On the same 1,000×15 BiRM pools, the PRMBench double-newline and Skywork single-newline rules robustly change 52.7%, 44.0%, and 63.2% of Qwen, Math-Shepherd, and Skywork winners (Figure 1). PRMBench yields one block for 14,454/15,000 candidates and every candidate in 816 pools, versus a median six Skywork blocks. This contrast measures whole-response against multi-block parsing on this fixed BiRM corpus; Appendix S gives all six rule pairs, filtering effects, and the input-identity control.

Source versus canonical changes 13.2–16.9% of winners, despite Kendall $\tau_b = 0.878$–0.914. Qwen and Math-Shepherd were frozen before outcomes; Skywork and the public-rule extension came later. All 18 parser-pair author final-accuracy intervals include zero.

Exhaustive and expert-annotated audits. On the common 8,000 synthetic traces, formally admissible partitions change verdicts for all three PRMs and enable conditional ranking reversals (Table 4). Our prespecified reducers are min for Qwen and Math-Shepherd and mean for Skywork; Qwen's released BoN evaluation instead uses a product of step probabilities (Zhang et al., 2025). ASR conditions on an incorrect candidate that initially trails its fixed correct comparator. The table also reports exact ordering certificates on 2,000 four-candidate pools. Only 15.9–42.6% of fine-policy winners strictly dominate every rival across independent admissible partition choices. Attainable wrong reversals identify possible orderings, rather than a score-only attack rate; ties are conservatively uncertified.

On ProcessBench-P, VFR/ASR@16 are 29.56%/5.22% for Qwen ($n = 230$ eligible pairs) and 11.94%/23.46% for Math-Shepherd ($n = 179$). Qwen PRMBench-P has 39.45% VFR and 5/126 conditional ranking successes. These results distinguish within-trace verdict instability from the ability to defeat an available correct comparator.

6 Separating aggregation and fresh-forward dependence

Changing boundaries alters both aggregation inputs and the serialized model context. Figure 2 isolates what the controls can establish. Synthetic scores change even at fixed block count. Holding the target block's start, end, and content fixed, an earlier boundary change alters its readout, while later boundary changes are near zero. Actual last aggregation also changes, although fixed-forward last is exactly invariant. An exploratory same-count Qwen-generated paragraph analysis also yields winner changes under the released product reducer (Appendix I). These results locate dependence before aggregation; the controls do not separate marker identity, position shifts, and attention.

Table 4: Exhaustive synthetic audit. Rates and individual 95% problem-cluster intervals are percentages. ASR@16 conditions on the stated eligible pairs; certification quantifies all partitions on 2,000 pools. Mechanism controls are separated into Figure 2.

A. Verdict flips and conditional ranking

Model	VFR [95% CI]	ASR@16 [95% CI]	Eligible
Qwen	16.0 [15.0,16.9]	18.0 [16.4,19.5]	2,852
Math-Shepherd	9.2 [8.4,9.9]	11.5 [10.5,12.7]	3,379
Skywork	71.1 [69.6,72.5]	55.4 [53.3,57.5]	3,660

B. Exact ordering over all legal partitions

Model	Fine certified	Correct certified	Wrong reversal
Qwen	42.6 [40.4,44.8]	42.2 [40.1,44.4]	57.1 [54.9,59.2]
Math-Shepherd	36.7 [34.6,38.8]	31.8 [29.8,33.8]	67.4 [65.4,69.4]
Skywork	15.9 [14.3,17.5]	15.9 [14.3,17.5]	84.2 [82.5,85.7]

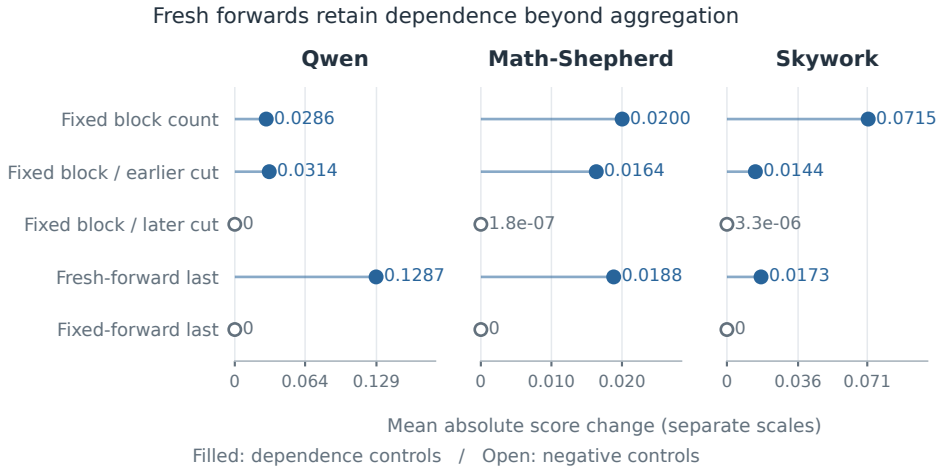

Figure 2: Mean absolute score changes on the 8,000 synthetic traces. Filled points show dependence controls; open points are negative controls. Fixed-count uses each PRM's prespecified study reducer; earlier/later cuts keep the target block fixed. Read each raw-score axis separately. These are descriptive magnitudes, not additive causal components.

Reducer choice still matters: Skywork VFR falls from 71.1% with mean to 37.8% with min, remaining above Qwen and Math-Shepherd under the same min reducer (16.0% and 9.2%). Appendix U reports all six model–reducer cells, calibration and eligibility, and fixed-forward diagnostics. The synthetic witness in Appendix O illustrates both components: its last score changes from 0.82894 to 0.39222 under a fresh forward, whereas endpoint-subset replay cannot change last.

7 Frozen-policy confirmation and transfer

In the earlier BiRM and MATH-500 studies, each candidate receives its maximum score among the first B frozen views; the highest-scoring candidate wins. View order is fine, fully merged, then hash-seeded partitions; B includes fine and shorter spaces are exhausted. The $B = 2$ policy therefore differs from merged-only scoring. The exploratory 500-pool BiRM study develops this policy; the following confirmations use separately frozen protocols.

Reserved-source confirmation. On 1,000 reserved questions, $B = 8$ changes 362 Qwen and 537 Math-Shepherd winners. Requiring both endpoint margins to exceed 4×10^{-5} leaves 25.4% [22.7, 28.1] and 43.9% [40.8, 47.0] robust changes, passing the frozen two-model 20% gate. Merged-only scoring also changes 51.9% [48.8, 55.0] and 47.5% [44.4, 50.6] robustly, so maxi-

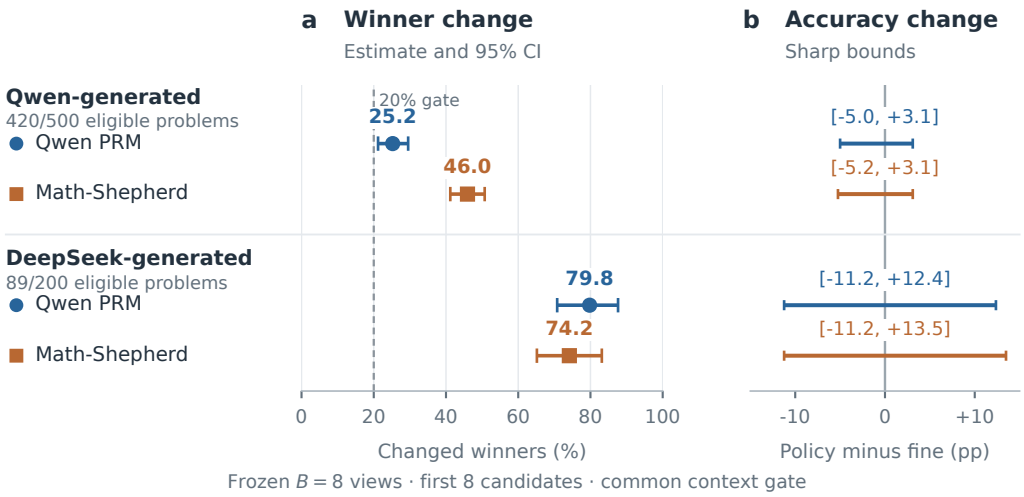

Figure 3: Frozen MATH-500 transfer on eight candidates from Qwen (420/500 eligible problems) and DeepSeek (89/200 sampled problems eligible) generators. (a) Robust $B = 8$ winner changes and individual 95% confidence intervals; all four lower endpoints exceed 20%. (b) Sharp accuracy-effect bounds retain unresolved verifier outcomes and all cross zero.

mizing across views is not necessary for decision changes. Figure 4 and Table 9 retain every frozen budget. Fresh GPU replay checks 59 views per model with zero maximum discrepancy.

Outcome metadata are joined after the score-only audit. Author-label accuracy changes are +0.6pp $[+0.0, +1.2]$ and +1.1pp $[+0.1, +2.2]$, but independent-verifier sharp bounds cross zero: $[-0.4, +1.5]$ and $[-0.3, +2.2]$pp. A gain-permutation baseline, which shuffles candidate score gains within each problem, is reported separately in Appendix F; it is not repeated-inference noise or a success gate.

Cross-dataset and cross-generator transfer. Figure 3 extends the same policy to the common-context MATH-500 cohorts. All four cell-specific lower endpoints exceed the frozen 20% gate, while every accuracy-effect bound crosses zero. Fresh GPU replay checks 60 views per PRM with zero discrepancy. Context eligibility remains consequential: 80/500 Qwen and 111/200 DeepSeek pools are excluded whole, and 300 benchmark problems were never sampled for DeepSeek generation. The frozen evidence is the new Math-Shepherd transfer and four-cell conjunction, conditional on these cohorts.

Additional views have different effects across cohorts (Figure 4). In the 89 retained DeepSeek pools, $B = 2, 4, 8$ select identical winners for both PRMs despite later score gains. Every candidate has eight distinct views and at least six atoms, so the plateau does not reflect lattice exhaustion. Later views do change winners in reserved BiRM and Qwen-generator pools. This is a descriptive cohort comparison, not a causal generator effect.

Decision instability and utility. The fresh harmful case shows that turnover can affect correctness. In the earlier cohorts, correct-to-correct and wrong-to-wrong replacements explain why identity turnover can coexist with little measured accuracy change (Appendix L). In the AI-qualified Qwen set, the direct join yields 12 correct-to-correct and one unknown-to-unknown replacement; the space-padded join yields 14 and one, respectively. A separate stratified 100-pair GSM8K judge audit reports 64 ties among 65 correct-to-correct replacements (Appendix M); these selected judgments do not establish a population utility effect.

8 Implications for reproducible scoring

A completed-trace scorer should declare normalization and base atomization, the partition rule and its owner, serialization, checkpoint, readout, reducer, tie-breaking, and runtime randomness. Ownership

affects the admissible choices: on the 84-pool diagnostic cohort, independent candidate partitions expand mean feasible winner-set size from 4.58 to 8.01 for Qwen and from 3.55 to 11.65 for Math-Shepherd, relative to a shared whole-pool rule (Table 7). Reporting the checkpoint alone cannot resolve this ambiguity.

Two executable contracts illustrate the tradeoff. A deterministic raw-text canonicalizer maps declared-equivalent inputs to one representation with one forward. A fixed multi-view contract freezes hash-seeded masks and a reducer: mean, 0.1 quantile, minimum, or mean minus population SD. The last is a dispersion penalty, not a confidence bound. With deterministic scoring, either specifies a reproducible reward; stochastic scoring additionally requires a sampling protocol and repeatability assessment. Appendix O provides the runnable canonicalizer and complete interface details.

These contracts provide reproducibility guarantees whose utility and computational cost can be evaluated separately. On the exploratory 500-pool cohort, $K = 8$ multi-view policies change 46.4–58.6% of Qwen and 24.8–28.0% of Math-Shepherd winners at about 7.67× the fine-policy token cost, without reliable accuracy gains (Table 16). Canonicalization guarantees the declared syntactic invariance; full-lattice certificates require exhaustive coverage. In the fresh study, independently reconstructing the baseline partition restores identical inputs for all 7,968 scored views; 16 repeated CPU controls also return identical scores. This closes the tested marker-placement channel and separates partition reproducibility from baseline utility.

9 Limitations

Our fixed-payload equivalence class formalizes syntactic admissibility; semantic-step qualification is assessed separately on reviewed subsets. Fresh-study annotation combines score-blind Codex judgments and arithmetic checks with subsequent PRM-score-hidden review of reversal cases by one independent human; non-reversal and excluded cases retain their initial qualification. The equal-count results apply to 28 text-eligible pools, with calibrated, case-score-blind AI review of all 224 pairs conducted after aggregate outcomes were known. Broader human qualification and alignment with Qwen's private training boundaries remain validation targets. The Level-5 stratum is label-motivated; its CPU audit covers every correctness change but not every identity change. Estimates describe the screened cohorts and stated policies: they do not establish systematic accuracy degradation, general adaptive-search superiority, or deployment prevalence. Earlier multi-model Qwen analyses use our min reducer; the equal-count and fresh studies use the released BoN product. Natural pools are context-gated, and finite-view audits cover the queried partitions. The stratified quality audit uses one closed judge, so independent-judge replication is needed to assess quality equivalence. PathFinder's mixed runtime builds and GenPRM's seed variation further delimit the auxiliary extensions (Appendices H and V).

10 Conclusion

PRM-PART shows that partition choice changes verdicts and candidate selection while normalized ordered payload is held fixed. Exact lattice certificates complement evidence from three open PRMs, two independently frozen cohorts, and two rollout generators. Fixed-forward and fresh-forward controls locate dependence partly before aggregation. Score-only boundary search also produces a CPU-confirmed correct-to-wrong full-pool selection, demonstrating that partition dependence can affect final-answer correctness. These findings motivate an explicit contract for normalization, base atomization, partition ownership, serialization, readout, reduction, tie-breaking, and runtime policy. The resulting framework supports further evaluation of semantic-step qualification and the stability of model and benchmark rankings.

Reproducibility statement

The `reproducibility/` supplement pins model revisions and protocols and provides portable CPU reanalyses of selected cached statistics, parser and sampler source, input hashes, and the exact witness. It includes the response-bearing 420-pool Qwen MATH-500 ledger, the complete 28-pool local FP32 equal-count experiment, all 224 three-model blinded reviews, and independent CPU forwards for robust flips. A separate model-card marker-join control provides full-pool input hashes and step

scores. The fresh study adds all scored candidate texts, per-step scores, frozen protocols, qualification records, 511 CPU audit receipts, and a weight-free replay of both policies and all budgets. The complete 16,384-generation archive is hash-indexed separately; it is not bundled. The earlier 32-pool semantic audit and three-seed GenPRM records remain included. For the blinded candidate-quality audit, the supplement provides the aggregate summary and SHA256 hashes of the pair manifest, rubric, and output schema; the 200 double-order closed-judge responses, three adjudications, and full input files are not redistributed in the package. The README maps each check to its inputs and exclusions; other full GPU caches, weights, training code, and local reproduction of the closed judge are unavailable.

AI USE STATEMENT

TraeCode and Codex assisted with code and document inspection, literature checks, feedback on experimental protocols and statistical interpretation, analysis implementation, numerical verification, and manuscript editing. The blinded GPT-6-Astra candidate-quality audit and three-model Qwen semantic-step review are AI judgments. The fresh study uses one score-blind Codex annotator for mathematical labels and unit completeness, assisted by recorded arithmetic checks; independent CPU inference separately verifies the scoring results. The audit methods and available review materials are reported. One independent human reviewer subsequently reassessed reversal cases with PRM scores hidden and agreed with the AI judgments on those cases. The authors remain responsible for the analyses, citations, and final claims.

REFERENCES

Lang Cao, Renhong Chen, Yingtian Zou, Chao Peng, Wu Ning, Huacong Xu, Qian Chen, Yuxian Wang, Peishuo Su, Mofan Peng, Zijie Chen, and Yitong Li. More bang for the buck: Process reward modeling with entropy-driven uncertainty. *arXiv preprint arXiv:2503.22233*, 2025. URL https://arxiv.org/abs/2503.22233.

Wenxiang Chen, Wei He, Zhiheng Xi, Honglin Guo, Boyang Hong, Jiazheng Zhang, Rui Zheng, Nijun Li, Tao Gui, Yun Li, Qi Zhang, and Xuanjing Huang. Better process supervision with bi-directional rewarding signals. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025. URL https://arxiv.org/abs/2503.04618.

Peixuan Han, Runhui Wang, Ketan Ramaneti, Jie Hao, Gerald Friedland, and Chris Kong. Cliff: Learning process rewards from the first mistake. *arXiv preprint arXiv:2609.02817*, 2026. URL https://arxiv.org/abs/2609.02817.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. *arXiv preprint arXiv:2103.03874*, 2021. URL https://arxiv.org/abs/2103.03874.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

Yuliang Liu, Junjie Lu, Zhaoling Chen, Chaofeng Qu, Jason Klein Liu, Chonghan Liu, Zefan Cai, Yunhui Xia, Li Zhao, and Jiang Bian. Adaptivestep: Automatically dividing reasoning step through model confidence. *arXiv preprint arXiv:2502.13943*, 2025. URL https://arxiv.org/abs/2502.13943.

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. *arXiv preprint arXiv:2505.19706*, 2025. URL https://arxiv.org/abs/2505.19706.

Qwen Team. Qwen2.5-Math-PRM-7B model card, 2025. URL https://huggingface.co/Qwen/Qwen2.5-Math-PRM-7B.

Ibne Farabi Shihab and Fariya Afrin. Quality-diversity stress tests for process reward models: What archive coverage can and cannot certify. *arXiv preprint arXiv:2608.08008*, 2026. URL https://arxiv.org/abs/2608.08008.

Ibne Farabi Shihab, Sanjeda Akter, and Anuj Sharma. Detecting proxy gaming in rl and llm alignment via evaluator stress tests. *arXiv preprint arXiv:2507.05619*, 2025. URL https://arxiv.org/abs/2507.05619.

Ibne Farabi Shihab, Fariya Afrin, Sanjeda Akter, and Anuj Sharma. Est-prm: Stress-testing process reward models before they become load-bearing. *arXiv preprint arXiv:2606.00437*, 2026. URL https://arxiv.org/abs/2606.00437.

Skywork AI. Skywork-o1-open-prm-qwen-2.5-7b model card. Hugging Face model repository, 2024. URL https://huggingface.co/Skywork/Skywork-o1-Open-PRM-Qwen-2.5-7B.

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*, 2025. URL https://aclanthology.org/2025.acl-long.1230/.

Rishabh Tiwari, Aditya Tomar, Udbhav Bamba, Monishwaran Maheswaran, Heng Yang, Michael W. Mahoney, Kurt Keutzer, and Amir Gholami. Reward under attack: Analyzing the robustness and hackability of process reward models. *arXiv preprint arXiv:2603.06621*, 2026. URL https://arxiv.org/abs/2603.06621.

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

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024. URL https://arxiv.org/abs/2312.08935.

Teng Wang, Zhangyi Jiang, Zhenqi He, Shenyang Tong, Wenhan Yang, Yanan Zheng, Zeyu Li, Zifan He, and Hailei Gong. Towards hierarchical multi-step reward models for enhanced reasoning in large language models. *arXiv preprint arXiv:2503.13551*, 2025. URL https://arxiv.org/abs/2503.13551.

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

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. URL https://arxiv.org/abs/2505.14391.

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. *arXiv preprint arXiv:2501.07301*, 2025. URL https://arxiv.org/abs/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. *arXiv preprint arXiv:2504.00891*, 2025. URL https://arxiv.org/abs/2504.00891.

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*, 2025. URL https://aclanthology.org/2025.acl-long.50/.

Chaoyang Zhou, Shunyu Liu, Zengmao Wang, Di Wang, Rong-Cheng Tu, Bo Du, and Dacheng Tao. Intra-trajectory consistency for reward modeling. *arXiv preprint arXiv:2506.09096*, 2025. URL https://arxiv.org/abs/2506.09096.

A Fresh wrong-selection experiment

Prospective scope and acquisition. The main protocol was fixed locally before inspecting new generation text or fresh-study PRM scores; it was not externally preregistered. From the pinned MATH test release, we select 1,024 questions from 3,293 eligible questions, disjoint by normalized prompt from the recorded exposure ledger. This does not exclude semantic paraphrases or pretraining exposure. Qwen2.5-Math-7B-Instruct generates 16 candidates per question using temperature 0.7, top-p 0.95, no top-k truncation, at most 3,000 new and 4,096 total tokens, and a separate frozen random stream per candidate. All 16,384 jobs are accounted for: 15,710 reach EOS and 674 reach the length limit. Length stops remain acquisition failures and are not shortened into completed solutions. Source and model revisions, prompts, seeds, and execution profiles accompany the supplement.

Pair qualification and stopping rule. Recover blank-line units while retaining every character; merge forward dangling colons, standalone headings, or unbalanced math delimiters, appending a final dangling fragment to the previous unit. The primary pair screen requires 6–20 complete units, at most 14,000 response characters, unambiguous answer extraction, and no reserved markers, code fences, LaTeX environments, or heading/list lines. Automated answer comparison only proposes candidates. In fixed SHA256 order, inspect at most four proposed-correct and two proposed-wrong candidates per question. Enroll the first fully sound derivation and first derivation with an identified mathematical error and wrong final answer, requiring all original units to be complete. Stop at 100 qualified pairs or the 1,024-question source cap, with no score-based replacements. Of 122 mechanically mixed questions, 76 qualify before the cap; both PRMs retain all 76 after token preflight. Initial qualification used one score-blind Codex annotator with recorded arithmetic/symbolic checks. A subsequent independent human check corroborates judgments on the reversal-case subset, as detailed below.

Subsequent independent human review. One independent human reviewer subsequently reassessed outcome-selected reversal cases with PRM scores hidden and agreed with the AI judgments on those cases. This provides independent corroboration within the reviewed subset; initial qualification records for non-reversal and excluded cases remain AI-based. We report the reviewer count, case-selection scope, score masking, and qualitative agreement. The supplement preserves the original AI annotations; case-level human-review records and a numerical agreement denominator are not included.

Interface, budget, and search. Every view groups the same units into exactly four nonempty contiguous blocks. The baseline minimizes squared deviations of cumulative character positions from quartiles, breaking ties lexicographically. Qwen uses its native chat template with `<extra_0>` directly after each block; Math-Shepherd uses its native question/response form and a space followed by the two-character Cyrillic marker (U+043A/U+0438). Exhaustive tokenizer preflight over all three-cut choices checks raw-text identity, the ordered non-marker token sequence, total token count, four markers, final marker absolute position, and the 4,096-token cap. No candidate is replaced after this gate. Fresh forwards use FP32, eager attention, one unpadded input at a time, with TF32, cache, and dropout disabled. Model revisions are Qwen `0610740060112df12585d00a1c5f4624d2f59051` and Math-Shepherd `45dc0a3c9ec699b645085c098ed38dc99fba4617`.

The scorer sees neither reference answers, labels, competitor scores, nor a success signal. Adaptive search samples an unseen legal ±1-unit boundary move around the current product-best view, falling back to all unseen views if needed. Random search samples uniformly without replacement from all unseen views. Independent hash-derived streams and the same baseline fix both policies. Budgets $B \in \{1, 2, 4, 8, 16\}$ count the baseline; exhausted spaces are retained. The product of four probabilities is primary for both PRMs (Qwen's released BoN reducer and an explicitly chosen common reducer for Math-Shepherd). The secondary minimum is evaluated at the *product-selected* view, never optimized separately. This differs from the earlier ASR@16 convention of 16 additional queries.

Pair endpoint. The correct comparator stays at baseline; only the wrong member is searched. The primary endpoint at $B = 16$ requires correct-minus-wrong $> 4 \times 10^{-5}$ initially and wrong-minus-correct $> 4 \times 10^{-5}$ after search. Qwen has 68 initially correct-leading and eight wrong-leading pairs; Math-Shepherd has 57 and 19, with no within-margin ties. Table 5 gives every primary policy result. The Qwen comparison has one adaptive-only success, three random-only successes, and six shared successes (−2.94pp, exact paired $p = .625$); Math-Shepherd has five, zero, and 14 (+8.77pp, $p = .0625$). These discordance tests are two-sided; no cross-model adaptive advantage is established. The supplement retains all budgets, zero-margin sensitivity, and secondary minimum endpoints.

Table 5: Fresh correct/wrong pairs at total budget $B = 16$. Denominators condition on a correct-leading baseline with gap $> 4 \times 10^{-5}$; successful wrong-leading outcomes exceed the same margin. Rates and exact 95% Clopper–Pearson intervals are percentages.

Model	Policy	Reversals	Rate (%)	95% CI
Qwen	Adaptive	7/68	10.29	[4.24, 20.07]
Qwen	Random	9/68	13.24	[6.23, 23.64]
Math-Shepherd	Adaptive	19/57	33.33	[21.40, 47.06]
Math-Shepherd	Random	14/57	24.56	[14.13, 37.76]

Fixed pools and separate Level-5 stratum. The original pool experiment uses the first 128 frozen question ordinals and candidate indices 0–7 without answer-label filtering. A minimum-hash rule fixes one owned candidate before text, labels, or scores are inspected. All eight candidates must complete and contain at least four complete units; the target needs at least six. Unlike pair qualification, this pool screen permits headed units and does not require a sound derivation for candidates with correct final answers. After 83 mechanical and 12 complete-unit exclusions, 33/128 pools remain (25.78% coverage). Their 264 final-answer labels comprise 236 correct, 28 wrong, and zero unknown; 24 pools are all-correct, two all-wrong, and seven mixed. Only two pools have a wrong owned target and a correct rival. This composition limits opportunities for harm without constituting a measured harm rate.

That answer composition motivates a separate public Level-5 stratum: the first 128 Level-5 questions with original ordinal at least 128, using the same first-eight and hash-target rules. Its selection and eligibility are frozen before any fresh-study PRM scoring, but the motivation uses original-sample labels. We therefore do not pool its rates with the original sample. After 79 mechanical and 11 complete-unit exclusions, 38/128 remain (29.69%). Their 304 labels comprise 259 correct, 45 wrong, and zero unknown; 25 pools are all-correct, two all-wrong, and 11 mixed, with four wrong-owned mixed pools. The two pool strata are disjoint. Pair/pool question overlap is recorded in the supplement; shared score receipts never allow a rival to use extra views. All seven rivals remain at baseline even when a rival was also searched for the pair experiment.

Table 6: Full eight-candidate selection at $B = 16$. "Harm" means correct-to-wrong; "repair" means wrong-to-correct. Both policies search only the hash-owned target. The Level-5 sample has label-informed motivation and is reported separately.

Stratum	Model	Policy	Changes	Harm	Repair	Correct (before → after)
Original ($n = 33$)	Qwen	Adaptive	12	0	0	31 → 31
Original ($n = 33$)	Qwen	Random	13	0	0	31 → 31
Original ($n = 33$)	Math-Shepherd	Adaptive	11	0	0	30 → 30
Original ($n = 33$)	Math-Shepherd	Random	11	0	0	30 → 30
Level 5 ($n = 38$)	Qwen	Adaptive	12	1	1	34 → 34
Level 5 ($n = 38$)	Qwen	Random	14	1	1	34 → 34
Level 5 ($n = 38$)	Math-Shepherd	Adaptive	7	0	1	31 → 32
Level 5 ($n = 38$)	Math-Shepherd	Random	8	0	1	31 → 32

Correctness results and uncertainty. Table 6 separates winner identity from final-answer correctness. Every original-stratum policy/model has zero harm and zero repair; a zero count of 0/33 has exact 95% interval $[0, 10.58]$%, not a guarantee of harmlessness. In Level 5, both Qwen policies harm question 371 and repair question 219, producing zero net accuracy change. Math-Shepherd

has no harm and repairs question 175 under both policies. Qwen's 1/38 harm rate is 2.63% with interval $[0.067, 13.81]$%, conditional on this stratum. Net repair-minus-harm intervals subtract simultaneous 97.5% binomial intervals, giving conservative 95% coverage: $[-12.43, +12.43]$pp in the original sample, $[-15.62, +15.62]$pp for Level-5 Qwen, and $[-10.86, +15.65]$pp for Level-5 Math-Shepherd. Every net interval includes zero.

Auditable harmful witness. For question 371, the correct intersection is

$$[100, 999] \cap [9^2, 9^3 - 1] \cap [11^2, 11^3 - 1] = [121, 728], \qquad \frac{728 - 121 + 1}{900} = \frac{152}{225}.$$

Six candidates give this answer and two give 233/300. The owned target (index 3, zero-based) incorrectly treats $999_9 = 819$ as the largest three-digit base-9 number; digit 9 is invalid in base 9. It also incorrectly calls 999_{11} the base-11 maximum. The baseline winner (index 7) has a sound derivation. The target's nine original units are unchanged: adaptive search moves cuts from $(2, 4, 6)$ to $(1, 2, 7)$ and random search selects $(1, 7, 8)$. Both use 16 total queries; selected query indices are 12 and 9, zero-based. Table 1 gives the products. CPU scoring preserves both winner identities and yields baseline top-two gap 0.000408273, adaptive gap 0.000804140, and random gap 0.000987763, all above the fixed margin. This example is selected for illustration after outcomes, whereas its pool, ownership, and policies were fixed before scoring. Complete units include headings; grouping them does not certify that every resulting block is a natural semantic step or matches private training boundaries.

Numerical verification and audit scope. An independent input routine retokenizes all 7,968 unique GPU views (5,758 original and 2,210 Level-5), verifies the invariants, and reconstructs identical baseline inputs under scorer-owned canonicalization. A separate raw-record reader, without importing the original analysis functions, agrees on 6,080 pair-policy-setting and 1,420 pool-policy-budget endpoints. Independent CPU FP32 inference covers 511 receipts: 192/224 for original Qwen/Math-Shepherd and 48/47 for Level 5. The original audit includes all 49 positive pair-policy events (16 Qwen, 33 Math-Shepherd), every $B = 16$ changed pool, and hash-selected negative and baseline controls. Audited reversal, baseline-eligibility, and winner identities all agree. Sixteen separate canonicalization-control forwards are exactly repeatable.

The Level-5 protocol requested CPU replay of every changed winner. The executed post-score audit is narrower: all four wrong-owned mixed pools plus every accuracy-changing pool, five pools per model, including all eight baselines and both selected target views. It confirms every observed correctness change, but does not cross-device validate the remaining identity-only changes; this is a deviation from the planned audit coverage. Scores are not universally bit-identical: original Qwen/Math-Shepherd have 3/27 audited views above the descriptive 10^{-5} numerical alert, and Level 5 has 0/4. Maximum step differences are $1.66 \times 10^{-5}/1.82 \times 10^{-5}$ originally and $7.93 \times 10^{-6}/1.66 \times 10^{-5}$ in Level 5. No disagreement is dropped and CPU results do not replace primary GPU rates. Together, invariant checks over all scored views and CPU confirmation of every observed correctness change substantiate the fixed-content selection results.

Portable records. The supplement contains the original protocols, all scored candidate inputs and probabilities, mathematical/complete-unit review records, pool exclusions, model profiles, all CPU receipts, and the exact witness token IDs. A standard-library replay checks record hashes, reconstructs adaptive and random trajectories, and recomputes every reported budget and reducer endpoint without weights. The full acquisition archive, including all 16,384 raw generations, is separately hash-indexed; cached replay does not regenerate those samples or rerun neural inference.

B Related-work intervention matrix

Work	Fixed-payload counterfactual audit	Fixed order	Model input changes	Partition class	Fresh/fixed control	Selection/contract
Reward Under Attack	No	Sometimes	Yes	Text edits	No	No / No
Evaluator Stress Tests	No	Sometimes	Yes	Audited edits	No	No / Partial
EST-PRM	No	Sometimes	Yes	Add/reorder steps	Mechanical analysis	No / Partial
AdaptiveStep	Partial	Yes	Not isolated	Confidence/readout	No	Downstream / Selector
EDU-PRM	Not evaluated	Yes	Yes	Entropy anchors	No	Downstream / Selector
HRM	No	Yes	Yes	Training merges	No	Downstream / No
PRM-PART	Yes	Yes	Yes	Contiguous blocks	Yes	Yes / Yes

The counterfactual-audit column asks whether multiple partitions of the same payload are compared under a frozen scorer. "Partial" credits AdaptiveStep's test-time rating-position comparisons and aggregate Bo64 evaluation, without casewise winner-transition audits or full-lattice certificates. Its readout comparisons do not isolate model-input changes. EDU-PRM segments generated branches using uncertainty anchors. Our distinction is the joint audit objective, rather than a claim that their segmentation must change the text. All PRM-PART variants preserve normalized ordered payload; markers and readouts alter model inputs. Full-text audits are hash-pinned.

C Proof details

Proposition 1. For every legal contiguous partition, the visible endpoint set $E(\mathcal{P})$ contains m; conversely, cutting after precisely the indices in $E \setminus \{m\}$ realizes every subset $E \subseteq \{1, \dots, m\}$ containing m. The fully merged partition therefore returns the singleton sequence (r_m). Comparing it with the partition whose endpoint sequence is (r_i, r_m) gives $r_i = r_m$ for the mean, $r_m \le r_i$ for the minimum, and $r_i r_m = r_m$ for the product. Repeating the comparison for every $i < m$ proves necessity. Substitution into an arbitrary endpoint subset proves sufficiency. Last always returns r_m. For positive probability scores, the product equality reduces to $r_i = 1$ for every earlier endpoint; if zero is allowed, $r_m = 0$ is the degenerate alternative in which every product is zero.

Contextual last. Let $q_m^{\mathcal{P}}$ denote the final-block readout at atom m under partition $\mathcal{P}$. Under last aggregation, the actual forward score is exactly $q_m^{\mathcal{P}}$. It is constant over the declared equivalence class if and only if $q_m^{\mathcal{P}} = q_m^{\mathcal{P}'}$ for all legal $\mathcal{P}, \mathcal{P}'$. The fixed-forward replay holds the endpoint values fixed and hence removes this contextual term; the observed gap between actual and fixed-forward last in Figure 2 therefore occurs before aggregation.

Scorer-owned canonicalization. Suppose deterministic C maps declared-equivalent serializations to one representation, and H is deterministic. For legal $\mathcal{P}, \mathcal{P}'$, this gives $C(\tau, \mathcal{P}) = C(\tau, \mathcal{P}')$. Applying the same deterministic H to both sides yields $H(C(\tau, \mathcal{P})) = H(C(\tau, \mathcal{P}'))$. The conclusion is only as broad as the declared domain of C; it neither identifies semantic steps nor establishes utility.

D Ownership-aware utility and finite-query certification

Partition ownership and feasible winner sets. We reanalyze the completion-hashed 84×16 natural cohort without a new PRM forward. The shared family enumerates every complete whole-pool `SEARCH_VIEW_k`, scorer-owned canonical, and available pinned-public-parser profile. A public profile is incomplete on five questions (one missing candidate each), and `SEARCH_VIEW_4–8` are incomplete on one question (four candidates each); each incomplete vector is excluded as a whole rather than score-filled. The product family deduplicates only within model, candidate, and partition mask. Candidate-local scoring, batch size one, common frozen model profiles, and absence of cross-candidate normalization, context, or state make this product executable only under candidate-level boundary ownership.

Formally, $\Omega_{\text{shared}} = \{(P_g(c_1), \dots, P_g(c_n)) : g \in G\}$ while $\Omega_{\text{product}} = \prod_i \Pi_i^{\text{obs}}$. When every shared output is recorded in the product family, $\Omega_{\text{shared}} \subseteq \Omega_{\text{product}}$ and $W_{\text{shared}} \subseteq W_{\text{product}}$; no missing parser–candidate score is filled. Mean $|W|$ is 4.58 versus 8.01 for Qwen and 3.55 versus 11.65 for Math-Shepherd. Confirmed utility flips are $0/84$ versus $1/84$ and $1/84$ versus $4/84$. Among the six pools with a verified-wrong candidate, all possible winners are verified correct in $4/6$ versus $4/6$ Qwen pools and $3/6$ versus $1/6$ Math-Shepherd pools.

Table 7: Ownership-aware decision and utility identification. Intervals are source-problem bootstraps. "Nontrivial correct" conditions on the six pools containing a verified-wrong candidate; it counts pools for which every possible winner is verified correct.

PRM	Feasible domain	Mean $\|W\|$ [95% CI]	Necessary winner ($\|W\|$=1)	Confirmed flip	Possible-only flip	Nontrivial correct
Qwen	Shared profile	4.58 [4.32,4.85]	1/84	0/84	12/84	4/6
Qwen	Independent product	8.01 [7.69,8.37]	0/84	1/84	14/84	4/6
Math-Shepherd	Shared profile	3.55 [3.29,3.82]	4/84	1/84	13/84	3/6
Math-Shepherd	Independent product	11.65 [10.85,12.44]	1/84	4/84	17/84	1/6

Unknown labels and witnesses. One unknown candidate receives one shared binary completion across every strategy and partition. For a completion y, define $V(y, \Omega) = \max_{p\in\Omega} y_{w(p)} - \min_{p\in\Omega} y_{w(p)}$ and report $V_{\text{lower}} = \min_y V(y, \Omega)$ and $V_{\text{upper}} = \max_y V(y, \Omega)$. A unique unknown winner therefore has unknown utility level but zero partition-induced variability. The product analysis finds five correct/wrong PRM–problem events on four source problems; shared profiles realize one of them. Each exported winner has a complete 16-candidate partition combination and a strict stored top-two margin exceeding 8×10^{-5}. The threshold is inherited sensitivity, not a newly proved uniform error bound. Labels were inspected to choose the opposite-utility combinations, so these are offline, label-aware existence witnesses, not label-blind attack rates or deployment probabilities.

The population statistics above are zero-forward cache reanalysis. After freezing a separate label-free replay manifest, we additionally rerun all 118 unique candidate views forming the 12 selected correct/wrong whole-pool combinations (10 product and 2 shared) on two V100 workers. All fresh aggregate scores exactly match cache, all 12 winners are unchanged, and all margins remain above 8×10^{-5}. Labels are joined only by an independent CPU audit after GPU scoring. This targeted replay validates witness numerics; it is not a new 84-problem frequency estimate.

Strict versus tie-aware synthetic certification. The strict Math-Shepherd count underlying 36.70% is 734/2,000. Replaying the same scores with the deployed deterministic candidate-ID tie-break yields 738 necessary winners: four additional pools have an exact equality between the fine winner's lower endpoint and a rival's upper endpoint. Thus 734 is the conservative strict-dominance estimand and 738 is tie-break-aware, not a data, tolerance, or version discrepancy.

Finite-query boundary. Let $\Omega_{\text{full}} \subseteq \prod_{i=1}^{n} \Pi_i$ be the full admissible partition domain under the declared ownership rule, and define its attainable score vectors by

$$\mathcal{S}(\Omega) = \{(S(c_1, P_1), \ldots, S(c_n, P_n)) : (P_1, \ldots, P_n) \in \Omega\}.$$

For a score-vector set $\mathcal{T}$, let $W(\mathcal{T})$ contain all winners obtained from vectors in $\mathcal{T}$ using the fixed tie rule. If

$$\mathcal{S}_{\text{in}} \subseteq \mathcal{S}(\Omega_{\text{full}}) \subseteq \mathcal{S}_{\text{out}}, \quad \text{then} \quad W(\mathcal{S}_{\text{in}}) \subseteq W(\mathcal{S}(\Omega_{\text{full}})) \subseteq W(\mathcal{S}_{\text{out}}).$$

Write these winner sets as $W_{\text{in}}, W_{\text{full}}, W_{\text{out}}$, respectively. The inner vectors must come from executable admissible profiles; the outer set is a bound in score space, not an enlargement of the legal partition domain. Two differently labeled executable winners in W_{in} prove full-domain vulnerability. A common label for all candidates in W_{out} proves utility invariance. In contrast, single-label W_{in} is not a safety certificate, and a mixed-label outer set is not a vulnerability witness. Finite-query safety can fail by two black-box completions that agree on every query but assign an unqueried legal wrong-candidate point either a losing or winning value, provided the score range, tie rule, structural constraints, and domain coupling permit both. It can still be certified without exhaustion when valid global bounds, structural identities, saturation/tie priority, restrictive coupling, or a label-homogeneous pool rule out the unsafe completion. Here the generic probability outer box $[0, 1]^{16}$ certifies only the 60 all-verified-correct pools/model; observed extrema are not promoted to outer bounds. This is a standard indistinguishability boundary specialized to partition ownership, not a new general query-complexity theorem.

E Natural-study validation curves

Table 8 retains every frozen natural-pool search budget and every candidate-pool size from the native-context MATH-500 primary policy. It also reports a metadata-only composition replay: eligibility identifiers are reconstructed from the completion-hashed paired rows and joined to the pinned MATH-500 revision without reading answer labels. The nested eligible populations contain 445, 420, 401, and 367 problems at $N = 4, 8, 16, 32$. The shrinking coverage and differential retention show why the $N = 32$ result is a ranking replication on an eligible subset rather than a population accuracy estimate. An exact tokenizer census further shows that native-context full-population $N = 8$ is infeasible: 80/500 Qwen-generator pools contain an overlength candidate for at least one pinned PRM, and none becomes common-context feasible after full merging. Truncation would change the payload, while sliding-window or chunk aggregation would define a different scorer. Assigning every unscored problem either outcome yields full-500 robust winner-change bounds of $[21.2, 37.2]$% for Qwen

and $[38.6, 54.6]$% for Math-Shepherd; corresponding accuracy-effect bounds are $[-20.2, +18.6]$pp and $[-20.4, +18.6]$pp.

Table 8: Frozen natural-study validation curves (Panel A: Qwen2.5-Math-PRM on the exploratory 500×15 BiRM cohort; Panel B: Qwen2.5-Math-PRM on Qwen-generator MATH-500). Accuracy effects are policy minus clean; unresolved selected-candidate labels remain unknown, and selecting the same unknown candidate on both sides contributes exactly zero. Panel B uses independent score-only coordinate ascent per candidate with $Q = 32$ additional queries, at most 33 views including fine, and early termination at legal-space exhaustion. It is a different policy from the frozen $B \leq 8$ view list. Exact ties in this MATH-500 protocol select the earliest candidate index. Panel A intervals are paired 95% bootstrap intervals; Panel B intervals are sharp sample identification bounds.

A. Natural 500×15 search-budget curve

Budget	Score raised	Winner Δ	Accuracy Δ [95% CI]	Exhausted spaces
$B = 2$	1,186/7,500	25/500	-0.6 [-1.4,+0.0] pp	50/7,500
$B = 4$	3,721/7,500	152/500	-0.2 [-1.2,+0.6] pp	601/7,500
$B = 8$	5,162/7,500	185/500	+0.0 [-1.0,+1.0] pp	1,911/7,500

B. Native-context MATH-500 replication across candidate-pool size

N	Eligible	$Q = 32$ search winner Δ	C→W / W→C / unresolved	Accuracy-effect bounds
4	445/500	238/445 (53.5%)	1/3/17	[-2.25,+3.37] pp
8	420/500	296/420 (70.5%)	3/3/19	[-3.81,+3.10] pp
16	401/500	304/401 (75.8%)	0/5/26	[-3.74,+5.99] pp
32	367/500	314/367 (85.6%)	1/3/30	[-5.99,+5.45] pp

C. $N = 32$ native-context subset composition

Stratum	Retained	Total	Retention
Level 5	68	134	50.7%
Asymptote in problem	19	42	45.2%
Intermediate Algebra	52	97	53.6%
Geometry	23	41	56.1%

F Reserved-source score-only confirmation

Table 9 reports every frozen search budget and the deterministic fully merged contract on the 1,000-group confirmation cohort. "Robust" requires top-two margins above 4×10^{-5} at both clean and searched endpoints. The primary gate uses only $B = 8$ and requires both model-specific 95% interval lower endpoints to exceed 20%; merged and $B = 2, 4$ are mandatory sensitivity reports, not alternative gates. The gain-permutation baseline rates are 65.5% $[63.3, 67.6]$ for Qwen and 62.8% $[60.4, 65.3]$ for Math-Shepherd. For each problem, let f_i be candidate i's fine score and $g_i = \max_{j \leq K_i} s_{ij} - f_i$ its maximum observed gain. Independently permute the scalar gains across candidates, score candidate i by $f_i + g_{\pi(i)}$, and select the largest score with lexicographically largest trace-ID tie-breaking. Repeat 10,000 times using NumPy `default_rng` seeds 202609163 (Qwen) and 202610163 (Math-Shepherd), processing problems and trace IDs in sorted order. The reported brackets are the 2.5th/97.5th percentiles over permutations, not sampling confidence intervals. It is neither a token-level forward-noise model nor evidence against the observed partition dependence.

Table 9: Frozen score-only confirmation. Deterministic Δ uses the frozen tie-break; robust Δ additionally excludes numerically close endpoint decisions. Intervals are 10,000-replicate problem-bootstrap percentile intervals.

Model	Policy	Deterministic Δ	Robust Δ [95% CI]	Exact top ties
Qwen2.5-Math-PRM	Merged	645/1000	519/1000 [48.8,55.0]%	35
Qwen2.5-Math-PRM	$B = 2$	49/1000	39/1000 [2.7,5.1]%	64
Qwen2.5-Math-PRM	$B = 4$	291/1000	208/1000 [18.3,23.4]%	42
Qwen2.5-Math-PRM	$B = 8$	362/1000	254/1000 [22.7,28.1]%	56
Math-Shepherd	Merged	570/1000	475/1000 [44.4,50.6]%	59
Math-Shepherd	$B = 2$	545/1000	452/1000 [42.2,48.2]%	59
Math-Shepherd	$B = 4$	547/1000	453/1000 [42.2,48.3]%	49
Math-Shepherd	$B = 8$	537/1000	439/1000 [40.8,47.0]%	55

Only after the score-only analysis, two independent CPU audits, and both fresh-model GPU replays passed did we join outcomes. At $B = 8$, the author-label transitions are 2 correct-to-wrong versus 8 wrong-to-correct for Qwen and 9 versus 20 for Math-Shepherd. Independent-label and consensus sharp bounds retain unresolved selected candidates rather than imputing them; the cross-zero bounds are reported in the underlying completion-hashed analysis, whose full cache is not included in this supplement. The confirmation supports a reward-specification claim, not a selected-policy recommendation.

G Untouched canonicalization and matched payload control

A second confirmation freezes 1,000 source groups from 1,128 identifiers outside prior fit, calibration, evaluation, and method-development allocations. Before opening their responses, we fix `text-canonical-v1`, both PRMs, all 15 candidates, the tie rule, 10,000 problem-bootstrap draws, and a ±1pp practical-effect margin; no group may be replaced. An allocation audit verifies zero overlap with prior opened groups. All 1,000 groups pass the common native-context gate. Fresh scoring uses 30,000 policy forwards per PRM plus 16 deterministic replays with zero discrepancy.

Canonicalization changes 184/1,000 Qwen and 167/1,000 Math-Shepherd winners (147 and 132 robust; Table 10). Both author-label accuracy intervals lie inside the frozen $[-1, +1]$pp practical-equivalence margin. Independent-label sharp bounds, $[-0.9, +1.0]$pp and $[-0.9, +1.1]$pp, and wider consensus bounds do not identify equivalence across label streams. Author process-validity intervals also cross zero, overall and among changed winners; independent process annotations are unavailable. Qwen's author future-soft difference is +0.24pp $[+0.01, +0.52]$ overall, a descriptive result outside the success gate.

Table 10: Untouched source-group confirmation and matched content-changing control. Accuracy uses source author labels; the frozen ±1pp gate is described in text.

PRM	Policy	Winner Δ	Robust Δ [95% CI]	Accuracy Δ [95% CI] pp
Qwen	Canonical	184/1000	147/1000 [12.5,16.9]%	+0.3 [+0.0,+0.7]
Qwen	Confidence prefix	553/1000	389/1000 [36.0,42.0]%	+0.0 [-0.7,+0.7]
Math-Shepherd	Canonical	167/1000	132/1000 [11.1,15.3]%	+0.1 [-0.5,+0.7]
Math-Shepherd	Confidence prefix	393/1000	303/1000 [27.5,33.1]%	-0.7 [-1.6,+0.2]

As a payload-changing scale control, we add EST-PRM's deterministic "Clearly, " prefix to each source-fine block on the same candidates. It changes 553 Qwen and 393 Math-Shepherd winners (389 and 303 robust); delimiter-only robust-change rates are lower by 24.2pp $[20.8, 27.8]$ and 17.1pp $[13.7, 20.4]$. The control matches models, serializer, reducer, and context gate, but not EST-PRM's seven-phrase mixture.

H Post-outcome model and domain extensions

Hierarchical-verifier extension. PathFinder-PRM-7B (Pala et al., 2025) adds an error-aware hierarchical scoring interface, although its backbone is also Qwen-based. We apply its native two-stage scorer to the first 150 reserved-cohort pools (2,250 candidates). On identical candidates and masks, fine versus fully merged changes 64.0% of Qwen winners (95% CI [56.0, 71.3]), 55.3% of Math-Shepherd winners [47.3, 63.3], and 54.0% of PathFinder winners [46.0, 62.0]. Fine versus max over $B = 8$ views changes 32.7%, 50.0%, and 53.3%, respectively. Author-label accuracy ranges across these policies are 88.7–90.0%, 85.3–88.0%, and 88.7–89.3%. Fully merged means one block, not `text-canonical-v1`. PathFinder's interval overlaps both comparators; this extends the scoring-interface audit without establishing a cross-architecture sensitivity ranking.

The following extensions fix their complete analysis before inspecting their new scores, but the candidate outcomes already existed; neither is an untouched confirmation. Skywork uses its pinned official newline-flag serializer, scalar value head, and published mean aggregation. MATH-500 compares generator blank-line paragraph boundaries with content-identical `text-canonical-v1`; one of 89 DeepSeek-generator pools is rejected whole because its raw view exceeds native context. Every unresolved outcome remains in sharp bounds.

Table 11: Post-outcome third-model and raw/canonical transfer extensions. MATH rows compare generator blank-line boundaries against `text-canonical-v1`; Accuracy Δ bounds are canonical minus raw and retain unknown selected outcomes. MATH entries show eligible-cohort bounds, while full-500 bounds are wider.

Setting	PRM / policy	n	Robust winner Δ [95% CI]	Accuracy Δ bounds (pp)
BiRM	Skywork / canonical	1,000	169/1,000 [14.7,19.2]%	[-0.7,+1.1]
BiRM	Skywork / confidence	1,000	435/1,000 [40.5,46.6]%	[-0.8,+0.5]
Qwen generator	Qwen PRM / raw vs canonical	420	126/420 [25.7,34.5]%	[-3.6,+5.5]
Qwen generator	Math-Shepherd / raw vs canonical	420	166/420 [34.8,44.3]%	[-3.3,+1.0]
DeepSeek generator	Qwen PRM / raw vs canonical	88	58/88 [55.7,76.1]%	[-10.2,+10.2]
DeepSeek generator	Math-Shepherd / raw vs canonical	88	39/88 [34.1,54.5]%	[-11.4,+9.1]

I Native Qwen same-count diagnostic

We reanalyze completed per-step receipts for the 420 context-eligible MATH-500 problems generated by Qwen, each with eight fixed candidates. The baseline splits each generated response at its recorded blank-line paragraph boundaries; the alternative uses the existing deterministic `text-canonical-v1` blocks. Both are passed through the pinned Qwen chat template and `<extra_0>` readout. Within-block whitespace is collapsed in both views, and their normalized ordered payloads match. The input ledger has SHA256 prefix `7e86a76e`; this analysis makes no new model calls.

To control the direct effect of changing the number of scores, the alternative replaces the baseline only for the 123/3,360 candidates whose two block lists differ but have the same count (at least two). These candidates occur in 75/420 pools; every other candidate retains its baseline view. This is a deterministic candidate-level fallback diagnostic, not a shared parser applied uniformly to every response. Exact ties use the study's largest-trace-ID rule. Qwen's released BoN product is computed as a sum of log step probabilities for ranking (Zhang et al., 2025); the study's min reducer is a secondary check.

The product changes 21/420 winners, with no exact top-two tie at either changed endpoint. Min changes 32/420, of which 20 have top-two gaps greater than 4×10^{-5} under both policies. Under product, the 21 author-label transitions comprise 17 correct-to-correct, two wrong-to-wrong, one unknown-to-unknown, and one unknown-to-correct; no known correct-to-wrong transition is identified. Existing outcomes and earlier view scores were known when this diagnostic was specified, so these are descriptive counts without a new confirmation interval. The equality of block counts removes the factor-count explanation for the product comparison, while changed marker locations and model context remain part of the intervention.

This first same-count comparison covers the declared syntactic class; the separate rubric audit below targets semantically coherent adjacent cuts. The complete Qwen response and readout ledger

now permits portable CPU replay of this product comparison, alongside the compact counts in `reproducibility/native_qwen_same_count.json`.

Adjacent same-count boundary shifts (exploratory). We further reanalyze the completion-hashed Qwen L4 cache without a new forward pass. A score-blind text rule searches the recorded views of each candidate for two partitions with at least three equal-count steps that differ by moving one boundary across one adjacent canonical atom. It rejects obvious unfinished fragments, then favors the largest step count and breaks ties by mask order. The rule selects 377/3,360 candidates in 167/420 pools; every selected pair preserves ordered content, step count, and even the total native input-token count. No pool has all eight candidates selected, so unselected candidates use the same canonical fine view in both arms. This is a candidate-level fallback comparison, not a uniformly applied parser.

Under Qwen's released product reducer, 44/420 pools change winner; 28 changes have top-two log-product gaps above 4×10^{-5} in both arms and one involves an exact top-two tie. The study min reducer gives 35/420 changes. Product's author-label transitions are 40 correct-to-correct, two unknown-to-correct, and two unknown-to-unknown, with no identified known correct-to-wrong replacement. For one four-step, 438-token MATH-500 rational-root answer, moving a single boundary changes its log-product from -0.007806 to -0.002858; a fixed rival scores -0.005262, and both selected answers are author-labeled correct. This is a post-outcome existence illustration, not a rate among semantically qualified alternatives.

Case inspection finds awkward cuts that pass the text rule, so the 44/420 result remains a syntactic mechanism diagnostic. The subsequent score-blind rubric audit separately qualifies semantic steps. The complete Qwen response and readout ledger now permits portable CPU replay of the 44/420 product result; compact counts remain in `reproducibility/native_qwen_local_shift.json`.

Score-blind semantic-step audit (exploratory). We then froze a separate rubric before opening case-level scores: each block must complete a coherent local reasoning move, and neither boundary may leave an isolated heading, unfinished lead-in, or detached list item. A trace-ID hash selected 60 of the 377 text-screened pairs from distinct problems; a single Agent reviewed both full block lists with scores, winner changes, and answer labels hidden. It accepted 32 pairs and conservatively rejected 28. Every accepted pair preserves the ordered payload, step count, and total Qwen input-token count, and uses the native `<extra_0>` serializer. Only the accepted candidate changes view; all other candidates retain the same fine view in both arms.

With Qwen's released product reducer, four of the 32 affected eight-candidate pools change winner (equivalently four of all 420 pools under this sparse, frozen fallback policy). Two changes have top-two log-product gaps above 4×10^{-5} at both endpoints. All four replacements are author-labeled correct-to-correct, so this measures identity instability, not demonstrated accuracy harm. A new local FP32 forward pass on all candidates in the 32 affected pools reproduces the four CUDA FP16 cached flips with identical winner pairs; 31/32 pools have identical winner pairs across the two runtimes. Repeating each of the four changed candidates' two FP32 views three times gives identical step scores. For example, in the four-step, 302-token coefficient-sum solution to $g(x) = f(x-1)$, moving a complete inference about $g(1)$ between its neighboring blocks changes one candidate's local log product from -0.001376 to -0.002442, crossing a fixed rival at -0.001582. Both winners are correct.

Apple MPS FP16 produced nonfinite scores; MPS BF16 produced seven affected-pool flips but large per-step deviations from the CUDA cache, so neither local precision supplies the primary count. The four-case FP32/CUDA agreement supports a narrower, rubric-qualified semantic-step result, not a prevalence estimate over all coherent partitions or proof of alignment with Qwen's undisclosed training-step distribution. The initial qualification used one score-blind Agent; the later reversal-case human check does not validate the full qualification sample. Earlier aggregate and cached-score analyses were known when this audit was designed. The full blinded packet, judgments, 32-pool step scores, and CPU replay are included with the supplementary source; its README gives the exact commands.

J Prose-screened equal-count whole-pool audit

The previous diagnostics change the view of an eligible candidate while other candidates retain a fixed view. To compare two complete rerankers, we froze a new text-only rule before new local model forwards. Starting from the same 420 eight-candidate Qwen-generator MATH-500 pools, it requires every candidate's canonical atoms to read as complete prose moves: at least five alphabetic words, no leading markup bullet or heading, detached display formula, or unfinished connective, and a terminal punctuation mark or completed display formula. A pool is included only when all eight candidates meet this rule and each has two admissible adjacent-boundary shifts. This retains 28/420 pools and 224 candidates; no score or answer label enters selection.

For a candidate with n canonical atoms, each view merges exactly one neighboring atom pair, leaving $n - 1$ steps. View A removes the boundary after atom i, and view B removes the next boundary; the shifted whole atom belongs to a complete block in both views. We select i by the smallest trace-ID-and-position SHA256 digest among text-qualified choices. All 224 pairs keep the same ordered payload, number of native `<extra_0>` readouts, and total input-token count. Removing the marker tokens from the token-ID streams leaves exactly two local non-marker token-ID changes per pair, induced by tokenizer context around the moved boundary. A score-blind hash sample of 24 candidates from distinct retained problems was reviewed using the earlier semantic-step rubric; one Agent accepted both complete step lists in all 24. The frozen protocol and packet are included in the supplementary data.

We made 448 new local Qwen FP32 forwards, one per candidate and view, using the pinned checkpoint, declared space-padded marker serializer, and the released product of step probabilities. Seventeen of 28 complete eight-candidate pools change winner; 15 changes have both top-two log-product gaps above 4×10^{-5} and none is an exact tie. A separate reload and repeat of all 448 forwards reproduced all 3,594 step scores exactly and the same 17 winner changes. Sixteen transitions are author-labeled correct-to-correct and one is unknown-to-unknown. The result is conditional on the 28 prose-screened pools, rather than a frequency over all 420 pools or all coherent decompositions. The portable supplement contains every input, mask, step score, blind judgment, repeat score, and a CPU replay of the full-pool ranking.

To review complete rerankers rather than a sample, three Codex models (GPT-6-Sol, GPT-6-Luna, and GPT-6-Astra) separately inspected the two full step lists for every one of the 224 candidates, with Qwen scores, winner status, answer labels, and trace IDs hidden. Eight positive/negative controls in both arm orders gave each model 16/16 before its judgments were used. The initial 96-pair packet covered a conservative subset; the remaining 128 pairs were frozen and reviewed afterward. The reviewers' intersection accepts 215/224 pairs. Requiring all eight pairs in a pool to pass all three leaves 25/28 complete pools, with 15 product-policy winner changes, 13 above the fixed two-margin gate. The three excluded pools contain, respectively, incomplete `\frac` expressions, isolated numbered-list markers, and a dependent equality clause split into a separate block. A deterministic TeX scan independently finds the malformed expression. Fourteen of the 15 replacements are author-labeled correct-to-correct and one unknown-to-unknown. The reviewers were blind to case-level scores, but the extended review was conducted after the aggregate 28-pool score result was known; Astra was added after the first two reviewers accepted all 96 initial pairs. All 13 margin-robust flip pools were independently rerun under CPU FP32 (208 candidate-view forwards); all 26 winner identities and their margin gates match the prior MPS FP32 run.

As a serializer sensitivity defined after the spaced-join scores were known, we repeat all 25 fully reviewed pools with the direct `<extra_0>` join shown in Qwen's model card (Qwen Team, 2025), holding checkpoint, steps, candidates, precision, and product reducer fixed. This gives 13 winner changes, 11 with both gaps above the original 4×10^{-5} gate; 10 also clear a stricter descriptive 10^{-4} cut. Among 19 complete pools whose eight candidate pairs still have equal input-token counts under this join, 9 change winner and 7 are margin-robust. Independent CPU FP32 reruns all 11 robust-flip pools plus two additional pools (208 forwards); all winner identities and robust-flip margin gates agree with MPS, with maximum per-step difference 4.1×10^{-6}. All 13 changes are author-labeled correct-to-correct (12) or unknown-to-unknown (1). The join changes some individual winners; this is a protocol sensitivity on the same selected cohort, not a held-out estimate. The supplement includes every new input hash and step score.

As a stricter format sensitivity specified after case inspection, excluding every pool with a numbered item or dash-led phrase leaves 12 of the original 28. One has the incomplete fraction expressions, leaving 11 fully reviewed pools, four changes, and three with robust margins. For example, in MATH-500 counting-and-probability problem 339, the winning candidate changes from index 6 to index 2; its top-two log-product gaps are 1.63×10^{-4} and 9.30×10^{-5} in the two views. The portable supplement contains the blinded packets, controls, decisions, scanner report, complete eight-candidate scores, and CPU replay.

K Cross-dataset transfer details

Table 12 reports every frozen policy in the MATH-500 transfer. "Robust" again requires both endpoint margins above 4×10^{-5}. The primary gate uses $B = 8$ only; merged and $B = 2, 4$ are mandatory sensitivity reports. All four $B = 8$ interval lower endpoints exceed 20%, whereas every accuracy-effect sharp bound crosses zero.

Table 12: Frozen MATH-500 transfer across two rollout generators and two PRMs. Intervals use 10,000 problem-bootstrap draws; ties are searched-endpoint exact top ties.

Generator	PRM	Policy	Deterministic Δ	Robust Δ [95% CI]	Ties
Qwen generator	Qwen PRM	Merged	316/420	112/420 [22.6,31.0]%	5
Qwen generator	Qwen PRM	$B = 2$	281/420	94/420 [18.6,26.4]%	8
Qwen generator	Qwen PRM	$B = 4$	297/420	105/420 [21.0,29.0]%	5
Qwen generator	Qwen PRM	$B = 8$	303/420	106/420 [21.2,29.5]%	5
Qwen generator	Math-Shepherd	Merged	249/420	219/420 [47.4,56.9]%	13
Qwen generator	Math-Shepherd	$B = 2$	227/420	198/420 [42.4,51.9]%	13
Qwen generator	Math-Shepherd	$B = 4$	223/420	197/420 [42.1,51.7]%	13
Qwen generator	Math-Shepherd	$B = 8$	222/420	193/420 [41.2,50.7]%	11
DeepSeek generator	Qwen PRM	Merged	73/89	71/89 [70.8,87.6]%	0
DeepSeek generator	Qwen PRM	$B = 2$	73/89	71/89 [70.8,87.6]%	0
DeepSeek generator	Qwen PRM	$B = 4$	73/89	71/89 [70.8,87.6]%	0
DeepSeek generator	Qwen PRM	$B = 8$	73/89	71/89 [70.8,87.6]%	0
DeepSeek generator	Math-Shepherd	Merged	70/89	66/89 [65.2,83.1]%	3
DeepSeek generator	Math-Shepherd	$B = 2$	70/89	66/89 [65.2,83.1]%	3
DeepSeek generator	Math-Shepherd	$B = 4$	70/89	66/89 [64.0,83.1]%	3
DeepSeek generator	Math-Shepherd	$B = 8$	70/89	66/89 [65.2,83.1]%	3

L Outcome-diverse utility identifiability

Only six pools in the matched 84-problem cohort contain both verified-correct and verified-wrong candidates; conditional harm denominators are five for Qwen and four for Math-Shepherd. Identity turnover there carries little information about utility.

In a retrospective Qwen pilot over 32 frozen candidates per problem, label-free score eligibility followed by outcome filtering (≥ 2 verified-correct, ≥ 2 verified-wrong, $\leq 5\%$ unknown) retains 37/367 native-context pools. An eight-forward policy changes 25/37 winners, with one harmful versus two beneficial transitions among 33 source-correct opportunities: harmful ASR is 3.03% (exact 95% interval $[0.08, 15.76]\%$). The accuracy difference is +2.70pp $[-5.41, +13.51]$pp. The sole harmful event has zero searched margin; a post-hoc 4×10^{-5} tolerance leaves 0/21 harmful flips. Canonicalization prevents none of the observed harmful events.

Before new Math-Shepherd scores, we froze 78 relaxed pools (including all 37 strict pools), reachable branches, and the tolerance. Its 19,968-call replay changes 15/37 strict winners: 0 harmful among 29 opportunities (exact 95% upper endpoint 11.94%), one beneficial, and +2.70pp $[0.00, +8.11]$pp accuracy difference. In relaxed pools, harm is 0/63 (exact upper endpoint 5.69%) with paired bounds $[-2.56, +2.56]$pp. Canonicalization has no harmful event to prevent and changes four searched-correct strict winners to wrong versus two searched-wrong to correct. This outcome-selected cohort is too small to establish equivalence.

PRM	Policy	Fine acc.	Policy acc.	Δ [95% CI] (pp)	Harm	Benefit
Qwen	Merged	86.83	88.32	+1.50 [-1.20,+4.19]	7/290	12/44
Qwen	$B = 8$	86.83	88.62	+1.80 [+0.00,+3.59]	2/290	8/44
Math-Shepherd	Merged	73.35	75.15	+1.80 [-1.20,+5.09]	11/245	17/89
Math-Shepherd	$B = 8$	73.35	75.75	+2.40 [-0.60,+5.39]	8/245	16/89

Table 13: Post-hoc reanalysis on 334 strict mixed-outcome BiRM pools. Accuracy is in percent; Δ is policy minus fine in percentage points. Harm counts wrong policy winners among correct fine winners; benefit counts correct policy winners among wrong fine winners. All 5,010 candidate labels in this subset are determinate. Both policies use the frozen lexicographically largest trace-ID tie rule. Intervals use 10,000 problem bootstrap draws (seed 202609201).

In the unfiltered 500-problem Qwen-32 population, 170 pools are verified mixed (sharp bounds 34.0–52.0% with unknown labels); only 56 satisfy the strict outcome rule (bounds 11.2–43.0%). A powered utility study needs 400–500 strict pools with outcome-relevant unknowns adjudicated blind to scores. Math-Shepherd passed context checks and 16 exact replays.

L.1 Post-hoc reanalysis of the reserved BiRM cohort

Post-hoc filtering of the reserved 1,000×15 BiRM cohort for strict mixed author-label outcomes (≥ 2 correct, ≥ 2 wrong, $\leq 5\%$ unknown) yields 334 pools. We compare fine, fully merged (one block, distinct from `text-canonical-v1`), and max-over-B=8 policies. This outcome-selected subset falls short of the predeclared 400–500-pool blinded gate.

The searched policy has 2 harmful versus 8 beneficial transitions for Qwen and 8 versus 16 for Math-Shepherd; merged-only has 7 versus 12 and 11 versus 17. Math-Shepherd's searched accuracy effect is +2.40pp $[-0.60, +5.39]$pp. These descriptive, outcome-filtered results do not establish utility equivalence.

M Blinded candidate-quality boundary audit

This post-outcome process-quality audit froze 100 fine-versus-merged winner pairs from 64 label-complete GSM8K problems before judge calls: 65 correct-to-correct, five correct-to-wrong, seven wrong-to-correct, 13 wrong-to-wrong, and 10 identical controls. Transition strata were deliberately selected, and some problems contribute multiple scale–PRM cells; Table 14 is descriptive.

For each pair, the prompt includes the problem, reference final answer, and solutions A and B, while hiding generator scale, PRM identity, partition, and which solution was selected by each policy. The frozen rubric instructs the judge to ignore verbosity, formatting, and prose polish and to compare intermediate correctness, logical completeness, unsupported jumps, and self-consistency; final-answer correctness is reported separately. We use the closed GPT-6-Astra endpoint (service label `gpt-6-astra`, unified provider, low reasoning effort; queried 2026-09-22) with a strict JSON schema. Each pair is presented in both A/B orders using a frozen random assignment. All 200 initial judgments are schema-valid, and 97/100 pairs map to the same outcome after undoing the order swap. The three disagreements all occur in the wrong-to-wrong stratum; a third blinded presentation adjudicates them. No initial response is `UNCERTAIN`.

The judge selects the correct-side solution for all 12 cross-correctness controls and ties all 10 identical controls. Among 65 correct-to-correct replacements, it reports 64 ties and one merged preference. One closed judge on selected pools cannot establish quality equivalence or a population harm rate; it finds no systematic process-quality disadvantage in this stratum.

The frozen pair manifest, rubric, and schema have SHA256 prefixes `dcbafc49`, `8a88e1d7`, and `0f6b4112`, respectively. The portable package includes the aggregate summary and these hashes; raw double-order responses, the three adjudications, and full input files are not included. The same-family 0.5B–7B generator analysis is exploratory: 64 common-label problems show no identified monotone accuracy-harm trend, but this sample is underpowered for one- or two-percentage-point effects and is not used to claim a scaling law.

Winner-label transition	Pairs	Fine better	Merged better	Quality tie
Correct → correct	65	0	1	64
Correct → wrong	5	5	0	0
Wrong → correct	7	0	7	0
Wrong → wrong	13	4	2	7
Identical-candidate control	10	0	0	10

Table 14: Final blinded pairwise quality judgments after third-presentation adjudication for the three initial order disagreements. Transition labels refer to fine-policy and fully-merged winner correctness, respectively. The stratified pairs are not a random sample from all winner changes.

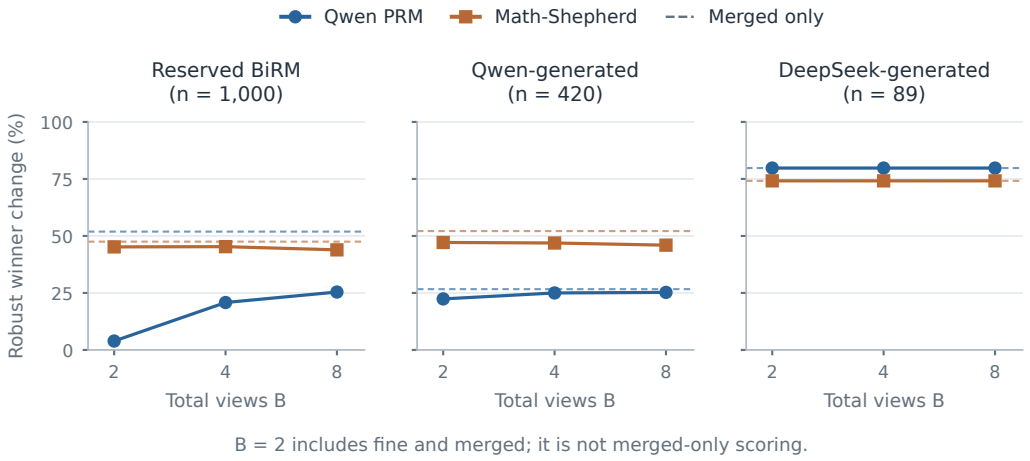

Figure 4: All frozen max-over-view budgets on the reserved BiRM and two MATH-500 cohorts. Points are robust winner-change rates relative to fine; intervals are tabulated in Tables 9 and 12. Dashed horizontal lines show the separate merged-only policy. $B = 2$ takes the candidate-wise maximum of fine and merged scores. Winner-change rates need not grow monotonically with B.

N Partition-source decomposition

Table 15 reports the post-hoc, exploratory paired decomposition of the completed L3/L4 score ledgers. It never subtracts cumulative winner-change rates to infer contribution: each adjacent nested-budget transition is classified at the source-problem level as a new departure from fine, retention of the same non-fine winner, return to fine, or switch between non-fine winners. The archived source-decomposition summary in the supplement additionally reports incremental candidate and pool-top score gains, C→W/W→C/C→C/W→W/unknown states, exact and near ties, token length, original unit count, score saturation, and legal-space exhaustion. All length, margin, saturation, and exhaustion bins are retained without significance-based selection.

Table 15: Exploratory source decomposition. Merged Δ is relative to fine; adjacent columns count any winner transition, including returns and switches. $B8 \neq B2$ directly measures selection information beyond the fine-plus-merged candidate maximum.

Cohort / PRM	n	Merged Δ	$B2 \to B4$	$B4 \to B8$	$B8 \neq B2$
Reserved / Math-Shepherd	1000	570	116	99	154
Reserved / Qwen	1000	645	267	202	334
Qwen gen. / Math-Shepherd	420	249	38	28	49
DeepSeek gen. / Math-Shepherd	89	70	0	0	0
Qwen gen. / Qwen	420	316	44	42	64
DeepSeek gen. / Qwen	89	73	0	0	0

O Implementation algorithms and minimal example

Public parser implementations. The pinned Qwen/PRMBench wrapper iterates over caller-provided steps and appends `<extra_0>` to each; the pinned Math-Shepherd wrapper analogously appends U+043A/U+0438. Separately, PRMBench's Best-of-N loader (commit `a1ae4eab`) splits a raw response on literal double newlines, strips parts, and discards empty parts. Skywork's official inference repository (commit `719b56b`) implements `response.split(step_token)` and its README calls it with a single newline. We exclude regex/action parsers that delete or rewrite textual `Step n:` prefixes because they violate the fixed normalized-payload condition. These sources establish concrete executable choices, not how often each rule is deployed.

Atomic units and frozen views. Synthetic atoms are the four-to-eight executable transitions emitted by the generator; ProcessBench/PRMBench atoms are their source expert steps. BiRM L3 units are the nonempty substrings before the source's whitespace-trimmed U+043A/U+0438 markers. MATH-500 L4 applies `text-canonical-v1`: collapse whitespace, then scan left-to-right; sentence punctuation or a completed display-math region closes a unit only outside paired brackets/math, with decimal, abbreviation, numbering-prefix, escape, and unclosed-region guards. Source spans are mapped back so concatenation reconstructs the raw response exactly. These deterministic syntactic units are not claimed to be semantic steps.

For m units, an $(m-1)$-bit mask encodes a legal contiguous partition: bit one cuts and bit zero merges. The exhaustive synthetic audit evaluates all 2^{m-1} masks. Its adaptive ASR@Q instead starts at fine and permits Q additional unique queries. Initialize Python `random.Random` from the first eight SHA256 digest bytes (big-endian integer) of UTF-8 `seed|trace_id|aggregator`, with seed 2025. At each query, choose uniformly among unseen single-boundary toggles of the current mask (candidate list ordered by boundary position); if none remains, choose uniformly from ascending unseen masks when the space has at most 4,096 masks, otherwise use `randrange(total)` with rejection of seen masks. Move only on strict score improvement, retaining the best score; stop issuing queries when all masks are seen. ASR@16 thus visits 8, 16, 17, 17, and 17 views for $m = 4, 5, 6, 7, 8$, respectively. Table 23 separates these coverage regimes.

Natural frozen-view studies use a different budget: B counts total views, including fine, with ordered unique masks $[2^{m-1}-1, 0, \ldots]$ capped at eight. For candidate c, $s_B(c) = \max_{j \le \min(B, K_c)} s(c, P_j)$; select by the pair $(s_B(c), \text{traceID}_c)$, largest first. Robust changes require both endpoint top-two gaps $> 4 \times 10^{-5}$. The synthetic ASR and legacy MATH-500 $Q = 32$ searches are adaptive; they must not be relabeled as this fixed-B policy.

```
UNITS(text): normalize whitespace; scan with math/bracket state;
emit only at guarded sentence punctuation or closed display math;
assert join(units)=normalized text.
VIEWS(units, trace_id): masks=[all-split, all-merge];
RNG=Random(int.from_bytes(SHA256(prefix|seed|trace_id), big));
repeatedly draw RNG.randrange(2^(m-1)), retaining only unseen masks
until min(8,2^(m-1)); return masks in that order.
SELECT(pool,B): for each candidate take max of the first min(B,K)
scores; return the largest (score,trace_id).
```

The view sampler uses the entire 32-byte SHA256 digest as a big-endian integer seed for Python `random.Random`; it samples integer masks with `randrange(total)`, rejecting duplicates while preserving acceptance order. L3 uses the UTF-8 string `evaluation-unseen|202609102|trace_id`; L4 uses `l4-math500-transfer|202609171|trace_id`. Fine and merged masks are deduplicated before sampling. The source bundle includes a standalone sampler and fixed-output examples. L3/L4 score-only bootstraps use 10,000 source-problem draws with seeds 202609161/202609172 and their declared model/cell offsets. New decompositions are post-hoc and descriptive.

Calibration and expert pairing. For each model–reducer pair, calibrate on 1,200 fine-view synthetic development traces. Candidate thresholds are midpoints of adjacent distinct scores plus `nextafter` values just below the minimum and above the maximum. Maximize the lexicographic tuple (balanced accuracy, ordinary accuracy, negative distance to the middle distinct score, negative threshold), classifying scores $\ge t_A$ as correct. Standardized scores use $(s - \mu_{\text{dev}})/\sigma_{\text{dev}}$ with population standard deviation; zero variance makes standardization unavailable.

ProcessBench grouping hashes UTF-8 question text after splitting and rejoining whitespace with one space: 3,400 traces form 2,842 clusters. Of these, 1,738 have no process-correct trace, 1,030 have exactly one, and 74 have more than one; only the exactly-one groups can supply a comparator. Their cluster sizes range from one to five, and singletons cannot supply an erroneous pair. Within eligible groups, retain each process-wrong trace whose fine score is strictly below the comparator: Qwen has 230 pairs and 12 successes, Math-Shepherd 179 and 42. PRMBench groups by the normalized original reference question stored by the converter, retaining distinct source rows rather than collapsing variants: 2,000 traces form 964 clusters (913 with zero correct, 48 with one, three with two). Qwen has 126 eligible pairs and five successes. These PRMBench reference-question clusters are a benchmark pairing convention; VFR remains the within-trace fixed-payload estimand. No answer-correctness inference is made from absent final-answer labels.

Serialization, readout, and scorer-owned policies. The PathFinder runs share one model revision but the runtime reported torch 2.7.1 for shard 1 and 2.13.0 for shards 2–3; shard 1 was not replayed on the upgraded build. Qwen uses the pinned official chat template and appends `<extra_0>` to every caller-supplied block; this study reads class-1 probability at each marker and takes the minimum. The Qwen release paper's BoN evaluation multiplies those step probabilities (Zhang et al., 2025); Appendix I applies that reducer to cached natural-response readouts. Math-Shepherd serializes the problem plus newline-separated blocks suffixed by U+043A/U+0438, takes $P(+ \mid +, -)$ at each marker, and takes the minimum. Skywork uses its pinned official `prepare_input`, newline reward flags, scalar readouts, and published mean aggregation. The model/profile records pair Qwen/Qwen2.5-Math-PRM-7B with `0610740`, peiyi9979/math-shepherd-mistral-7b-prm with `45dc0a3`, and Skywork/Skywork-o1-Open-PRM-Qwen-2.5-7B with `b89a895`; each scientific record stores masks, blocks, endpoint/readout positions, token-ID hashes, local scores, and completion receipts.

The full five-checkpoint mapping and adapter source hashes are in `reproducibility/model_protocols.json`. The extensions use `declare-lab/PathFinder-PRM-7B` at revision `84a7412` and `GenPRM/GenPRM-7B` at `0ea3fa7`; full 40-character revisions are recorded in that file. PathFinder's native search adapter predicts mathematical-correctness and consistency labels, returns -1 for a negative label, and otherwise conditions a second-pass correctness score on both positive labels (Pala et al., 2025). This differs from a single scalar readout, but does not remove step boundaries from its input.

For hidden scores $x_1, \ldots, x_K$, the evaluated reducers are mean, linearly interpolated empirical 0.1 quantile, observed minimum, and

$$\mathrm{LCB}_K = K^{-1} \sum_{j=1}^{K} x_j - \sqrt{K^{-1} \sum_{j=1}^{K} (x_j - \bar{x})^2}.$$

This coefficient-one population-SD LCB is a fixed conservative dispersion penalty, not a calibrated confidence bound. The raw-text canonicalizer removes only whitespace-delimited reserved protocol markers, normalizes whitespace, applies the scan above, and asserts that joining its blocks exactly reconstructs normalized content. It accepts no gold atom or label. A runnable minimal example is `PYTHONPATH=src python scripts/canonicalization_example.py`: both `Compute x. <extra_0> Return 3. <extra_0>` and its merged serialization map to blocks `[Compute x., Return 3.]`. The included parser source and example check identical content, spans, and blocks for these serializations.

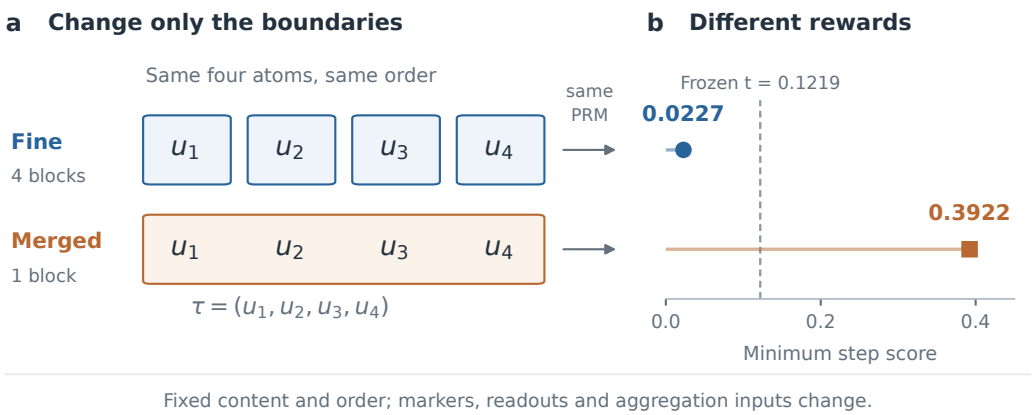

Figure 5: Partition dependence for the synthetic Qwen mechanism witness. (a) Both serializations preserve the same four ordered atoms. (b) The minimum step score changes from 0.0227 to 0.3922, crossing the frozen threshold 0.1219. Markers and model inputs differ; the normalized ordered reasoning payload is fixed.

Mechanism witness record.

The completion-verified trace is `synth-v7-candidate-test-000031-t0`. Its prompt and ordered atom strings are, verbatim:

```
Solve (-7/3)(x - 1) - 2/3 = 61/3 for x.

1. Start: (-7/3)(x - 1) - 2/3 = 61/3.
2. Expand the parentheses and collect constants: (-7/3)x + 5/3 = 61/3.
3. Subtract the left constant from both sides: (-7/3)x = 56/3.
4. Divide both sides by the coefficient of x: x = -8.
```

The system message is `Please reason step by step, and put your final answer within \boxed{}.` The user message is the prompt above. The assistant response joins the listed blocks with `<extra_0>` and appends `<extra_0>`; the merged response first joins the four atoms with one ASCII space, then appends one marker. Apply the pinned Qwen chat template with `tokenize=False`, `add_generation_prompt=False`, then encode the conversation. The marker is token 151651. Fine has 146 input tokens with zero-based readout positions 79, 105, 127, 143; merged has 140 tokens and readout position 137. Class-1 probabilities from float16 forwards with float32 softmax are $(0.99374664, 0.02271575, 0.67056787, 0.82893884)$ and (0.39222044), crossing $t_{\min} = 0.12187750$. The source bundle's `reproducibility/witness.json` preserves exact messages, token IDs, full-precision scores, and source/completion hashes.

Outcome verification. MATH-500 correctness uses the pinned verifier libraries `math_verify` and `latex2sympy2_extended`. An answer is Boolean only when there is exactly one parseable terminal `boxed{}` answer, no malformed/multiple box or substantive suffix, and verification completes; otherwise it is unknown and remains in the denominator. DeepSeek rollouts first isolate the frozen final-answer section. L3 independently parses the terminal source unit with the same pinned verifier and retains exceptions/non-Boolean outputs as unknown. Sharp sample bounds optimize all unknown binary completions while sharing the same unknown candidate across policy contrasts; bootstrap intervals are never substituted for these identification bounds.

P Auxiliary consistency result

A same-endpoint consistency loss reduces cross-partition KL by approximately 9–12% in a matched auxiliary study but does not improve candidate selection over ordinary multi-partition training with

the same ranking target. Lower disagreement is therefore not sufficient evidence of a useful scoring method, and EPC is not presented as a method contribution. This experiment and the associated optimizer diagnostics are auxiliary to the measurement paper.

Q Full scorer-owned natural-policy family

Table 16 reports every frozen K×reducer policy, rather than selecting the largest point estimate. All decisions use only cached scores and are written before labels are joined. The author-label intervals are paired, individual descriptive intervals. The underlying policy analysis retains sharp identification bounds for independent and conservative-consensus labels; the full policy cache is outside this supplement. The same K has the same forward/token cost across reducers because reduction is post-processing.

For Qwen, $K = 8$ mean changes 232/500 winners with a -0.2pp estimate ($[-1.2, +0.8]$), while $K = 8$ LCB changes 262/500 with 0.0pp ($[-1.0, +1.2]$). Every $K = 8$ reducer uses 57,496 forwards and 17,097,276 input tokens, or 7.67× the fine-policy token cost. The existing 500×15 raw-text run changes 97 Qwen and 86 Math-Shepherd winners; both author accuracy differences are -0.2pp with intervals crossing zero. The allocation-isolated 1,000×15 cohort is reported in Table 10. A paired difference-in-difference analysis does not identify a change in their comparison. These are post-outcome analyses, not a selected-policy claim.

Table 16: Complete post-hoc scorer-owned policy family on the natural 500×15 pool. Winner changes and transitions are relative to fine. Token cost is relative to fine. No row establishes a utility improvement after family-wide selection. For $K = 2$, the coefficient-one population-SD dispersion penalty (LCB) reduces algebraically to the observed minimum, which is why the $K = 2$ LCB and $K = 2$ observed-minimum rows coincide.

Model	Policy	Accuracy	Δ [95% CI] pp	Winner Δ	C→W / W→C	Token ×
Qwen	Fine	83.6%	–	–	–	1.000
Qwen	Merged	83.8%	+0.2 [-1.2,+1.6]	315	6/7	0.962
Qwen	K2 mean	82.8%	-0.8 [-1.6,-0.2]	208	4/0	1.962
Qwen	K2 q10	83.4%	-0.2 [-1.2,+0.8]	265	4/3	1.962
Qwen	K2 LCB	84.0%	+0.4 [-1.0,+1.8]	290	5/7	1.962
Qwen	K2 observed min	84.0%	+0.4 [-1.0,+1.8]	290	5/7	1.962
Qwen	K4 mean	83.6%	+0.0 [-1.0,+1.0]	233	3/3	3.917
Qwen	K4 q10	84.0%	+0.4 [-0.6,+1.6]	276	3/5	3.917
Qwen	K4 LCB	83.4%	-0.2 [-1.2,+0.8]	271	4/3	3.917
Qwen	K4 observed min	83.8%	+0.2 [-1.0,+1.4]	291	4/5	3.917
Qwen	K8 mean	83.4%	-0.2 [-1.2,+0.8]	232	4/3	7.667
Qwen	K8 q10	83.8%	+0.2 [-0.8,+1.2]	274	3/4	7.667
Qwen	K8 LCB	83.6%	+0.0 [-1.0,+1.2]	262	4/4	7.667
Qwen	K8 observed min	84.0%	+0.4 [-0.6,+1.6]	293	3/5	7.667
Math-Shepherd	Fine	78.6%	–	–	–	1.000
Math-Shepherd	Merged	78.0%	-0.6 [-2.8,+1.6]	278	16/13	0.956
Math-Shepherd	K2 mean	78.4%	-0.2 [-1.6,+1.2]	135	7/6	1.956
Math-Shepherd	K2 q10	78.2%	-0.4 [-1.6,+0.8]	74	6/4	1.956
Math-Shepherd	K2 LCB	78.0%	-0.6 [-1.6,+0.4]	46	5/2	1.956
Math-Shepherd	K2 observed min	78.0%	-0.6 [-1.6,+0.4]	46	5/2	1.956
Math-Shepherd	K4 mean	78.8%	+0.2 [-1.2,+1.6]	165	6/7	3.905
Math-Shepherd	K4 q10	78.8%	+0.2 [-1.0,+1.4]	125	4/5	3.905
Math-Shepherd	K4 LCB	78.0%	-0.6 [-1.8,+0.6]	131	6/3	3.905
Math-Shepherd	K4 observed min	78.6%	+0.0 [-1.2,+1.2]	110	5/5	3.905
Math-Shepherd	K8 mean	78.2%	-0.4 [-1.8,+1.0]	140	7/5	7.659
Math-Shepherd	K8 q10	78.6%	+0.0 [-1.2,+1.2]	124	5/5	7.659
Math-Shepherd	K8 LCB	78.4%	-0.2 [-1.4,+1.0]	124	5/4	7.659
Math-Shepherd	K8 observed min	78.6%	+0.0 [-1.2,+1.2]	139	5/5	7.659

R Boundary-rule ranking details

The post-outcome source/canonical ranking analysis included in Figure 1 was specified before computing the new ranking and stratum summaries, after the earlier winner and accuracy summaries were

known. It verifies all completion receipts, 1,000 pools of 15 unique candidates, exact payload reconstruction, and repeat parser determinism. Kendall's τ_b retains exact score ties. Pair agreement maps score differences of magnitude at most 4×10^{-5} to ties; decisive inversions condition on both rules being outside that tolerance. Top-k uses the fixed descending-score, then descending-trace-ID order and separately records near-tied kth boundaries. For the source-delimited versus `text-canonical-v1` comparison, robust winner changes remain 16.3/13.3/17.6% at zero tolerance and 6.5/11.1/13.4% at 10^{-3} for Qwen, Math-Shepherd, and Skywork, respectively.

All pre-specified length, source-step-count, and fine-margin bins are retained in the machine-readable report. Longer/more finely segmented pools are descriptively less stable: for example, robust top-1 disagreement rises from 10.7% to 18.8% between the shortest and longest Qwen quartiles and from 13.7% to 24.5% for Skywork. These are exploratory associations, not a causal length effect. Difficulty is absent from the frozen BiRM schema, so no correctness-derived proxy is introduced. Author process-validity transitions among changed winners are 9/14, 12/12, and 9/18 correct-to-wrong/wrong-to-correct; independent and consensus process labels are unavailable and remain unknown. The original source/canonical analysis retains further boundary-tie and stratum details; only the matched-pool records and selected summaries are included in this supplement.

S Public-parser matched-pool extension

This post-outcome extension froze parser rules and pairwise analysis before new scores, after candidate outcomes and earlier source/canonical results were known. All 15,000 candidates pass payload, order, and context checks. Exact input matches reuse verified scores; other views require 16,204 new forwards per model (48,612 total). All 48 deterministic replays match exactly.

Block-count boundary of the comparison. A pre-score block census (Table 17) finds one PRM-Bench block on 14,454/15,000 responses and all 15 candidates in 816/1,000 pools. Source and Skywork rules have a median of six blocks; Skywork exactly matches the source block list for 12,481 candidates. Thus the public-rule comparison on BiRM is chiefly coarse versus fine.

Rule	Median blocks	Mean blocks	One-block candidates	All-one-block pools
Source delimiters	6	6.54	2/15,000	0/1,000
Text canonical v1	5	5.81	518/15,000	1/1,000
PRMBench double newline	1	1.18	14,454/15,000	816/1,000
Skywork single newline	6	6.69	659/15,000	1/1,000

Table 17: Pre-score block census for the fixed BiRM cohort. Candidate counts use 15,000 responses; all-one-block pool counts require all 15 candidates to parse as one block. PRMBench and Skywork rules are cross-applied to this corpus, not sampled from deployment traffic. The compact hash-pinned census is in `reproducibility/parser_block_audit.json`.

Table 18: All six pairwise comparisons among source-delimited, `text-canonical-v1`, PRMBench double-newline, and the Skywork single-newline nonempty projection. Robust changes require both top-two margins to exceed 4×10^{-5}; intervals are 10,000-replicate source-problem bootstraps. Top-3/5 reports overlap fractions. Author accuracy Δ is right minus left. Qwen/Math-Shepherd source–canonical cells reuse the verified confirmation results in Table 10.

PRM	Rule pair	Robust top-1 Δ	τ_b	Pair agree	Decisive inv.	Top-3/5	Author Δ pp
Qwen	Source–Canonical	14.7 [12.5,16.9]%	0.878	93.6%	5.9%	88.0/90.5%	+0.3 [+0.0,+0.7]
Qwen	Source–PRMBench	52.7 [49.6,55.8]%	0.518	74.6%	23.6%	53.9/64.9%	-0.2 [-1.2,+0.8]
Qwen	Source–Skywork-NL	8.2 [6.6,9.9]%	0.908	95.3%	4.5%	91.2/93.3%	+0.2 [-0.3,+0.8]
Qwen	Canonical–PRMBench	53.4 [50.3,56.5]%	0.525	75.0%	23.3%	54.5/65.4%	-0.5 [-1.5,+0.5]
Qwen	Canonical–Skywork-NL	18.5 [16.1,20.9]%	0.832	91.3%	8.2%	84.0/87.3%	-0.1 [-0.6,+0.4]
Qwen	PRMBench–Skywork-NL	52.7 [49.6,55.8]%	0.521	74.7%	23.4%	53.2/64.7%	+0.4 [-0.6,+1.4]
Math-Shepherd	Source–Canonical	13.2 [11.1,15.3]%	0.914	95.3%	4.1%	89.6/92.9%	+0.1 [-0.5,+0.7]
Math-Shepherd	Source–PRMBench	43.3 [40.2,46.4]%	0.648	80.8%	17.2%	62.0/72.0%	+0.2 [-1.2,+1.6]
Math-Shepherd	Source–Skywork-NL	12.7 [10.7,14.8]%	0.940	96.8%	2.9%	92.0/94.9%	+0.3 [-0.4,+1.0]
Math-Shepherd	Canonical–PRMBench	45.5 [42.5,48.6]%	0.650	80.9%	17.1%	61.8/72.0%	+0.1 [-1.3,+1.5]
Math-Shepherd	Canonical–Skywork-NL	21.5 [18.9,24.1]%	0.879	93.5%	5.8%	84.3/89.9%	+0.2 [-0.7,+1.1]
Math-Shepherd	PRMBench–Skywork-NL	44.0 [40.8,47.1]%	0.650	80.9%	17.1%	61.8/72.2%	+0.1 [-1.3,+1.5]
Skywork	Source–Canonical	16.9 [14.6,19.3]%	0.883	94.1%	5.8%	88.6/91.5%	+0.1 [-0.2,+0.4]
Skywork	Source–PRMBench	62.5 [59.5,65.6]%	0.482	74.3%	25.9%	52.5/62.9%	+0.5 [-0.4,+1.4]
Skywork	Source–Skywork-NL	14.4 [12.2,16.5]%	0.915	95.7%	4.3%	91.0/93.7%	+0.0 [-0.5,+0.6]
Skywork	Canonical–PRMBench	60.4 [57.3,63.4]%	0.496	75.0%	25.2%	53.3/64.0%	+0.4 [-0.6,+1.4]
Skywork	Canonical–Skywork-NL	25.2 [22.6,27.9]%	0.834	91.7%	8.3%	83.9/88.1%	-0.1 [-0.8,+0.6]
Skywork	PRMBench–Skywork-NL	63.2 [60.2,66.1]%	0.482	74.3%	25.9%	51.7/62.6%	-0.5 [-1.5,+0.5]

Table 18 reports every pair, including ranking and author-accuracy differences; all 18 accuracy intervals include zero. Cached records retain tolerance, length, source-step-count, and reference-margin strata.

By pre-score PRMBench block count, the all-one-block pools contain 422/816, 351/816, and 517/816 robust changes for Qwen, Math-Shepherd, and Skywork; the other pools contain 105/184, 89/184, and 115/184. This is a descriptive stratum comparison.

An exact-input control finds 441 pools where all PRMBench views have one block and every Skywork view matches the source's native input tokens and reward flags. Source versus Skywork changes no winners there; PRMBench versus Skywork equals source versus PRMBench: 203, 164, and 284 robust changes for Qwen, Math-Shepherd, and Skywork. This post-outcome stratum and its receipts are in `reproducibility/public_input_equivalence.json`.

Table 19: Fixed-count candidate-set retention for every rule pair. Each cell is intersection divided by the common accepted count; no cell is a quality-loss estimate.

PRM	Rule pair	Top 10%	Top 25%	Top 50%
Qwen	Source–Canonical	89.1%	94.2%	97.7%
Qwen	Source–PRMBench	54.4%	76.5%	91.6%
Qwen	Source–Skywork-NL	97.9%	96.9%	98.1%
Qwen	Canonical–PRMBench	55.7%	76.9%	91.9%
Qwen	Canonical–Skywork-NL	89.3%	94.1%	97.3%
Qwen	PRMBench–Skywork-NL	53.9%	75.8%	91.3%
Math-Shepherd	Source–Canonical	94.1%	96.2%	98.2%
Math-Shepherd	Source–PRMBench	79.9%	84.7%	91.8%
Math-Shepherd	Source–Skywork-NL	96.8%	98.5%	98.4%
Math-Shepherd	Canonical–PRMBench	79.5%	85.5%	92.1%
Math-Shepherd	Canonical–Skywork-NL	94.5%	96.0%	98.0%
Math-Shepherd	PRMBench–Skywork-NL	79.1%	84.7%	91.7%
Skywork	Source–Canonical	93.0%	94.2%	97.8%
Skywork	Source–PRMBench	51.7%	64.3%	83.9%
Skywork	Source–Skywork-NL	97.2%	98.0%	98.5%
Skywork	Canonical–PRMBench	53.6%	66.1%	84.5%
Skywork	Canonical–Skywork-NL	91.9%	93.1%	96.8%
Skywork	PRMBench–Skywork-NL	52.3%	64.8%	84.0%

Skywork's literal split retains empty pieces on 548 candidates; cross-model comparison uses the ordered nonempty projection because the other interfaces require nonempty blocks. The pairwise rates describe this fixed corpus, not parser use in deployment.

T Fixed-acceptance filtering

We compare global top-10%, 25%, and 50% acceptance among the same 15,000 candidates under source and canonical scoring, fixing count and score/trace-ID order (Table 20). At 10%, 1,500 candidates are retained per policy; overlap is 89.1%, 94.1%, and 93.0% for Qwen, Math-Shepherd, and Skywork. A 4×10^{-5} cutoff sensitivity leaves 239, 163, and 205 candidates in the symmetric-difference membership gap.

Table 20: Fixed-count filtering under two boundary rules. Retained is intersection divided by each policy's accepted count; Jaccard uses the union. Author-label deltas are descriptive percentage points. Full-population independent final/process labels are unavailable, so no quality equivalence is claimed.

PRM	Accept	Retained	Jaccard	Changed problems	Δ correct pp	Δ all-valid pp
Qwen	10%	89.1%	80.4%	143/1000	-0.1	-0.1
Qwen	25%	94.2%	89.1%	204/1000	-0.0	-0.3
Qwen	50%	97.7%	95.5%	200/1000	+0.1	-0.1
Math-Shepherd	10%	94.1%	88.8%	91/1000	+0.1	-0.6
Math-Shepherd	25%	96.2%	92.7%	156/1000	+0.1	+0.1
Math-Shepherd	50%	98.2%	96.5%	172/1000	+0.1	+0.2
Skywork	10%	93.0%	86.9%	101/1000	-0.1	-0.3
Skywork	25%	94.2%	89.1%	212/1000	-0.1	-0.3
Skywork	50%	97.8%	95.8%	180/1000	+0.0	-0.2

Fractions and models were fixed before exporting author labels. Across nine cells, final-correctness composition changes range from −0.1 to +0.1pp and all-valid changes from −0.6 to +0.2pp. No

equivalence margin or independent full-population labels are available. The machine-readable output retains further composition, tie, and tolerance diagnostics.

U Reducer, search-coverage, and selection diagnostics

These post-feedback diagnostics reuse completion-checked scores; they add no model forwards.

Common-reducer comparison. Table 21 evaluates min and mean on the same 8,000 traces, retaining reducer-specific development thresholds. ASR@16 uses reducer-specific adaptive searches and eligibility; the final column uses common eligible pairs. Search trajectories can differ, whereas exhaustive VFR uses identical partitions. Skywork's min reducer lowers VFR by 33.29pp relative to

PRM	Reducer	t_A	VFR [95% CI]	ASR@16 [95% CI]	Eligible	Common ASR (n)
Qwen	min	0.121877	15.97 [15.02,16.91]	17.95 [16.39,19.50]	2,852	15.73 (2,740)
Qwen	mean	0.784330	19.95 [18.76,21.12]	17.82 [16.35,19.33]	3,754	9.27 (2,740)
Math-Shepherd	min	0.305700	9.18 [8.43,9.94]	11.54 [10.46,12.70]	3,379	11.03 (3,274)
Math-Shepherd	mean	0.566997	31.60 [30.01,33.19]	68.26 [66.26,70.21]	3,522	66.71 (3,274)
Skywork	min	0.100012	37.79 [36.52,39.09]	26.18 [24.56,27.83]	3,293	25.44 (3,227)
Skywork	mean	0.233688	71.08 [69.60,72.52]	55.36 [53.31,57.46]	3,660	53.64 (3,227)

Table 21: Common-reducer reanalysis. Rates and individual 95% problem-cluster intervals are percentages. Common ASR uses 2,740, 3,274, and 3,227 eligible pairs for Qwen, Math-Shepherd, and Skywork, respectively. No best reducer is selected.

mean; Math-Shepherd's mean raises ASR from 11.54% to 68.26% (11.03% to 66.71% on common eligibility). Aggregation affects cross-model comparisons.

PRM	Reducer	Fresh-forward $\lvert\Delta\rvert$	Fixed-forward $\lvert\Delta\rvert$	Residual $\lvert\Delta\rvert$
Qwen	min	0.02996	0.11507	0.10969
Qwen	mean	0.15101	0.12426	0.11027
Math-Shepherd	min	0.02211	0.00727	0.01755
Math-Shepherd	mean	0.10538	0.10296	0.01727
Skywork	min	0.02270	0.01332	0.01380
Skywork	mean	0.06806	0.06053	0.01637

Table 22: Matched mechanism diagnostics under both reducers. For each trace, the cached control averages absolute changes across its non-fine legal partitions relative to fine, then averages over traces. Fixed-forward selects the corresponding endpoints from fine-view readouts; residual is fresh-forward minus fixed-forward signed change before taking its absolute value. These magnitudes do not add and are not causal percentages.

Budget versus legal-space coverage. Table 23 separates exhausted lattices from partially searched ones at the same additional-query budget.

DeepSeek selection flow. Generation samples 200 of 500 IDs without replacement using `random.Random(2025).sample(sorted(ids),200)`. All have eight candidates; 111 fail the common context gate, leaving 89 (17.8% of the benchmark). Raw-paragraph versus canonical scoring excludes one further pool ($n = 88$).

Every retained candidate has eight unique masks; the minimum atom count is six, so even the smallest legal space has 32 partitions. None of the 712 retained lattices is exhausted at $B = 8$. The absence of winner transitions beyond $B = 2$ therefore describes ranking stability under these particular added views, not lack of remaining partitions.

V Exploratory generative-verifier audit

A GenPRM-7B extension (Zhao et al., 2025) uses a custom verification prompt, an isolated executor, 150 pools (2,250 candidates), and three seeds (20260920–20260922). Four of 13,500 expected

PRM	Atoms	Views/space	Success/eligible	ASR@16 [95% CI]
Qwen	4	8/8	63/658	9.6 [7.2,12.1]
Qwen	5	16/16	76/577	13.2 [10.3,16.3]
Qwen	6	17/32	104/543	19.2 [15.6,22.9]
Qwen	7	17/64	127/537	23.6 [19.7,27.8]
Qwen	8	17/128	142/537	26.4 [22.2,30.7]
Math-Shepherd	4	8/8	41/682	6.0 [4.2,7.9]
Math-Shepherd	5	16/16	81/684	11.8 [9.4,14.4]
Math-Shepherd	6	17/32	84/677	12.4 [10.0,15.0]
Math-Shepherd	7	17/64	77/669	11.5 [9.0,14.2]
Math-Shepherd	8	17/128	107/667	16.0 [13.4,18.9]
Skywork	4	8/8	257/747	34.4 [29.9,38.9]
Skywork	5	16/16	369/734	50.3 [45.6,55.1]
Skywork	6	17/32	420/714	58.8 [54.0,63.4]
Skywork	7	17/64	454/736	61.7 [57.2,66.1]
Skywork	8	17/128	526/729	72.2 [68.2,76.0]

Table 23: ASR@16 by atom count under each model's prespecified study reducer. The view count includes fine and therefore equals $\min(17, 2^{m-1})$. All strata and their eligible denominators are retained; intervals are 10,000 problem-cluster bootstrap draws (seed 202609201).

Statistic	Retained	Context-excluded
Problem pools / candidate traces	89 / 712	111 / 888
Mean fine-input tokens, Qwen tokenizer	2,013.7	3,747.3
Mean fine-input tokens, Math-Shepherd tokenizer	2,051.3	3,850.2
Median pool-maximum fine tokens, Qwen	2,698	4,524
Median candidate atom count	71	121
MATH levels 1 / 2 / 3 / 4 / 5	12 / 17 / 29 / 19 / 12	6 / 10 / 23 / 27 / 45

Table 24: Descriptive attrition census on the sampled DeepSeek population. Token counts include the prompt and native serialization for the fine view; eligibility checks every frozen view under both tokenizers. A pool is rejected if any candidate/view overflows, so the excluded group's mean need not exceed 4,096.

cells are missing. Strict $[0, 1]$ feasible-winner enumeration gives cross-view flip counts $[119, 120]$, $[119, 120]$, and 116 (77.3–80.0%); among 148 complete pools, the rates are 80.4%, 79.7%, and 77.7%. Within-view seed changes affect 138/148 pools, and paired bootstrap cross-view-minus-seed-noise intervals include zero ($[-10.8, +5.4]$ and $[-8.8, +7.4]$pp). The stored `canonical` key means fully merged. Seed variation under this custom prompt limits attribution to partition choice.

The supplement pins prompts, source hashes, one-sample inference (1,024 tokens, three-second code limit), and all 13,500 request seeds derived from SHA256 of `base:pool:trace:view`. Missing probabilities range jointly over $[0, 1]$ with persisted-order tie-breaking; no complete view has an exact top tie. Paired intervals use 20,000 problem-bootstrap draws on 148 complete pools. Missing-score bounds and bootstrap intervals address different uncertainties.