EST-PRM: Stress-Testing Process Reward Models Before They Become Load-Bearing

Anonymous ACL submission

Abstract

Process reward models (PRMs) are widely used in language-model training with dense step-level supervision. They assume PRM scores are stable proxies for step correctness under label-preserving transformations. These transformations change reasoning structure but preserve final answers. We argue this assumption is not well validated. Such transformations can change how PRM scores relate to correctness signals, leading to different failure modes across models.To address this gap, we introduce **EST-PRM**, a stress-testing framework for dense process rewards. It applies three transformations: (1) step inflation, (2) dependency-aware step reordering, and (3) confidence markers. A vulnerability decomposition is defined that separates reward inflation from loss of correctness sensitivity. Five PRM-style models are evaluated on 4,687 reasoning chains from MATH-500, GSM8K, and PRMBench.The results indicate clear differences in vulnerability patterns across models. Math-Shepherd shows the strongest sensitivity to position perturbations, with a Pearson correlation drop of 0.152 ± 0.038 and a $32.8 \pm 4.9\%$ score inflation rate. Qwen2.5-Math-PRM is most affected by step inflation, reaching a $47.6 \pm 4.3\%$ inflation rate. Confidence-based perturbations also distort reward calibration, revealing inconsistencies in correctness estimation. Three mitigation strategies are evaluated, highlighting trade-offs between robustness coverage and false-positive rates.

1 Introduction

A quiet methodological convergence is reshaping reinforcement learning for language models. Across a growing set of recent systems, including SDPO (Hübotter et al., 2026), G-OPD (Yang et al., 2026b), REOPOLD (Ko et al., 2026), Nemotron-Cascade 2 (Yang et al., 2026c), OpenClaw-RL (Wang et al., 2026), RLSD (Yang et al., 2026a), and SRPO (Li et al., 2026), researchers increasingly combine dense feedback, self-distillation, sample routing, and reward-based filtering to transform sparse correctness signals into stronger training and selection signals. This paradigm has yielded state-of-the-art performance in challenging reasoning settings. For instance, Nemotron-Cascade 2's 30B mixture-of-experts model (Fedus et al., 2022; Jiang et al., 2024) achieves gold-medal performance on the International Mathematical Olympiad and the International Olympiad in Informatics while activating only 3B parameters, and REOPOLD shows that a 7B student can match a 32B teacher while achieving $3.3\times$ faster inference. Despite differences in implementation, these systems rely on evaluators that assign credit at the level of intermediate reasoning steps.

Process reward models (PRMs) constitute one such class of evaluators. PRMs assign scores to intermediate reasoning steps rather than only to final outcomes. This design enables dense credit assignment over extended reasoning trajectories and makes PRMs effective for guiding search and optimization in reasoning models. However, their practical utility depends on a stronger assumption than agreement with ground-truth labels on naturally occurring traces: PRM scores must remain aligned with correctness under distributional shifts induced by policy optimization over high-reward trajectories. Under such selection pressure, any systematic deviation between PRM scores and true correctness becomes exploitable and introduces bias into both training and inference-time selection.

We argue that standard evaluations of PRMs are insufficient for deployment in such regimes. Existing benchmarks measure agreement with human or automatically verified labels on naturally generated reasoning chains (Lightman et al., 2024; Wang et al., 2024; He et al., 2025). While nec-

essary, these evaluations do not test robustness under label-preserving structural perturbations. This limitation aligns with prior findings in reward modeling for RLHF, where sparse outcome-based reward models show systematic vulnerabilities under stress testing (Amodei et al., 2016; Manheim and Garrabrant, 2018; Casper et al., 2023). However, dense step-level supervision expands the attack surface, since each intermediate scoring decision becomes a potential point of failure.

To interrogate this surface, we introduce three classes of label-preserving transformations that preserve final correctness while perturbing reasoning structure (Figure 1):

1. **Step-inflation** increases reasoning length through insertion of semantically redundant restatement or verification steps and alters the per-step reward distribution without affecting the final answer.

2. **Position-sensitivity** permutes steps under dependency constraints and exposes reliance on absolute step ordering rather than logical structure.

3. **Confidence-injection** augments steps with discourse markers (e.g., "clearly", "by definition") that remain semantically vacuous but may influence learned scoring heuristics.

Each transformation reflects a distinct hypothesis about structural failure modes in PRM scoring functions.

We conduct an empirical study on 4,687 retained reasoning chains across five PRM-style models with diverse architectures and training regimes: Math-Shepherd-7B (Wang et al., 2024), Qwen2.5-Math-PRM-7B (Zhang et al., 2025), RLHFlow PRM (RLHFlow Team, 2024), Skywork-o1-Open-PRM-8B (He et al., 2025), and DeepSeek-Math-7B-PRM (Shao et al., 2024). We also include a zero-shot Llama-3-8B-Instruct critic as a lightweight, format-insensitive baseline. Across models, we observe consistent but heterogeneous vulnerability patterns. Math-Shepherd and Skywork-PRM show strong sensitivity to step ordering, whereas Qwen2.5-Math-PRM and DeepSeek-PRM show stronger sensitivity to step-inflation. These results indicate that vulnerabilities vary across models and manifest as distinct dominant failure modes. Strong performance on natural traces does not therefore imply robustness under structural perturbations.

Our contributions include:

- A targeted attack suite for process reward models that evaluates robustness under label-preserving structural transformations and complements existing stress-testing frameworks for sparse reward models

- A formal vulnerability framework that separates reward inflation effects from degradation of correctness-aligned signal quality under structural perturbations

- A scalable experimental protocol over 1,000 problems (4,687 reasoning chains), including corrupted-chain validation and human verification to ensure label preservation

- A characterization of systematic, model-dependent failure modes across five PRM families, showing that robustness does not correlate straightforwardly with natural-input accuracy

- An evaluation of mitigation strategies with analysis of trade-offs between detection sensitivity and false-positive rates under adversarially perturbed reasoning traces

- A release of the full benchmark suite, attack implementations, and human-validated annotations to support reproducibility and robustness evaluation of PRMs in downstream applications

Our findings show that PRM robustness is highly sensitive to structural perturbations in reasoning traces, and that different models exhibit distinct dominant failure modes rather than uniform degradation. These results highlight the need for robustness evaluation beyond natural-input agreement and provide a foundation for more reliable deployment of process-level supervision.

2 Related Work

Our work builds on recent advances in process reward models, reward-model evaluation, and reward hacking analysis for learned evaluators. We examine the robustness of dense step-level supervision under label-preserving transformations. Below, we review the most closely related literature on PRMs while Appendix A presents a broader discussion of reward hacking and proxy

gaming,calibration methods, reinforcement learning pipelines, and alignment evaluation frameworks.

Process reward models, benchmarks, and EST-PRM positioning. Process reward models (PRMs) were introduced to enable dense, step-level supervision for reasoning tasks. Early open-source formulations such as those of Lightman et al. (2024) and Wang et al. (2024) established the basic paradigm of process-level scoring. Subsequent higher-capacity variants, including Skywork-PRM (He et al., 2025) and Qwen2.5-Math-PRM (Zhang et al., 2025), improved coverage and performance on mathematical reasoning benchmarks. Existing evaluation suites for PRMs mainly focus on detection of naturally occurring or synthetically injected reasoning errors at the step level and achieve strong baseline reliability under distribution shifts and adversarial error conditions. EST-PRM stands as complementary to this line of work: rather than evaluating whether a PRM identifies incorrect steps already present in the data, it tests whether PRM scores remain invariant under label-preserving transformations that do not alter correctness. Table 1 summarizes the relationship between EST-PRM and prior evaluation methodologies.

3 The EST-PRM Framework

A process reward model is a function $\mathrm{PRM}(q, s_1, \ldots, s_k) \in \mathbb{R}$ that scores a reasoning chain $(s_1, \ldots, s_k)$ for a question q. We write $x = (q, s_1, \ldots, s_k)$ for a scored chain and $Y(x) \in \{0, 1\}$ for its ground-truth correctness label. A stress transformation $\mathcal{A}$ maps x to a modified chain $\mathcal{A}(x)$ while preserving both the final answer and the ground-truth label, so that $Y(\mathcal{A}(x)) = Y(x)$. Because Y is by construction unchanged, any movement of the PRM score from $S_P(x)$ to $S_P(\mathcal{A}(x))$ is a movement that cannot be justified by correctness, and is therefore the quantity we want to measure.

EST-PRM measures two such movements. The first is reward inflation: the transformed chain receives a higher score even though its correctness label is unchanged. The second is correctness-signal degradation: across the benchmark, the correlation between PRM score and correctness weakens after the transformation is applied. These are not redundant. A PRM can inflate individual scores without losing population-level correlation (if inflation is consistent across both correct and incorrect chains), and it can lose correlation without any net inflation (if scores are merely shuffled within the existing range). Separating the two is the central diagnostic move of the framework.

Table 1: Comparison of PRM evaluation methodologies. EST-PRM focuses on dense-reward label-preserving gaming and reward inflation rather than natural-error detection alone.

Work	Unit tested	Perturbation	PRM specific?	Dense attack?
ProcessBench	Process steps	Natural errors	Yes	No
PRMBench	Process steps	Step-level errors	Yes	No
RewardBench	Outcome reward	Diverse heuristic/LLM errors	No	No
RLHF Hacking	Outcome reward	Adversarial inputs, mostly outcome-level	No	No
EST-PRM (ours)	Process steps	Step inflation, position shifts, confidence perturbations	Yes	Yes

Definition 1 (Correlation collapse)**.** Let $B = \{(x_j, y_j)\}_{j=1}^{n}$ be a benchmark with nonconstant labels, and let $S_P(x)$ be the scalar score assigned by PRM P to chain x. The correlation collapse of P under transformation $\mathcal{A}$ on B is

$$\Delta\rho(P, \mathcal{A}, B) = \rho\big(S_P(X), Y\big) - \rho\big(S_P(\mathcal{A}(X)), Y\big), \tag{1}$$

where ρ is the Pearson correlation over the empirical distribution induced by B.

Definition 2 (Score inflation and score-inflation rate)**.** The mean score inflation of P under $\mathcal{A}$ on B is

$$\Delta S(P, \mathcal{A}, B) = \frac{1}{n} \sum_{j=1}^{n} \left[S_P(\mathcal{A}(x_j)) - S_P(x_j) \right]. \tag{2}$$

For a threshold $\tau > 0$, the score-inflation rate is

$$I_\tau(P, \mathcal{A}, B) = \frac{1}{n} \sum_{j=1}^{n} \mathbf{1}\Big\{ S_P(\mathcal{A}(x_j)) > (1+\tau) S_P(x_j) \Big\}. \tag{3}$$

The main experiments use $\tau = 0.1$.

A concrete illustration of what each transformation does on a short example chain is given in Appendix F (Table 10), together with the full attack templates and dependency-graph constraints.

3.1 Formal Vulnerability Analysis

The empirical patterns we report in Section 5 are not arbitrary observations; each attack class is designed to exploit a structural property of common

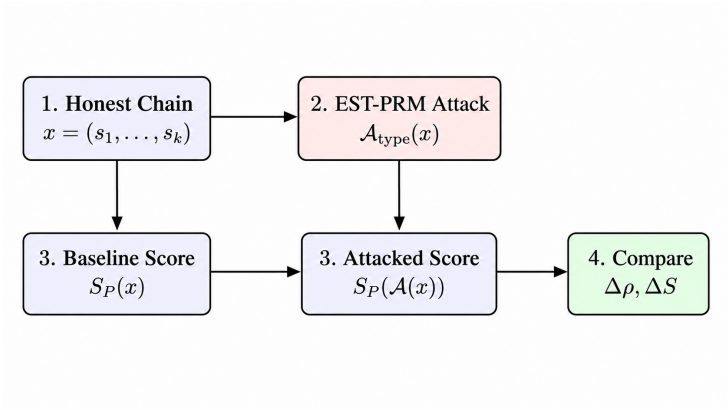

Figure 1: The EST-PRM pipeline. A question and its corresponding honest reasoning chain are subjected to a label-preserving attack transformation. Process reward model (PRM) scores are then evaluated before and after transformation to quantify correlation collapse and reward inflation, thereby isolating vulnerabilities arising from structurally induced reward gaming.

PRM scoring schemes. We make this connection precise here by isolating two such properties and giving conditions under which each attack class can or cannot move the score.

Definition 3 (Mean-aggregated PRM)**.** A PRM is mean-aggregated on a chain $(q, s_1, \ldots, s_k)$ if its scalar chain score can be written as

$$S_P(q, s_1, \ldots, s_k) = \frac{1}{k} \sum_{i=1}^{k} r_i, \tag{4}$$

where $r_i = \mathrm{PRM}_i(q, s_1, \ldots, s_i)$ is the reward assigned to the ith step under the model's native scoring format.

Definition 4 (Position-invariant PRM)**.** A PRM is position-invariant on a class of dependency-preserving permutations Π if

$$S_P(q, s_{\pi(1)}, \ldots, s_{\pi(k)}) = S_P(q, s_1, \ldots, s_k) \tag{5}$$

for every $\pi \in \Pi$. Failure of this condition means that at least one allowed permutation changes the score while preserving the multiset of reasoning steps and the final-answer label.

With these properties named, the effect of step-inflation on a mean-aggregated PRM follows directly from arithmetic.

Theorem 5 (Step-inflation under mean aggregation)**.** *Let P be mean-aggregated on a chain $(s_1, \ldots, s_k)$ with average step reward $\bar{r}$. Suppose a step-inflation transformation inserts m additional restatement steps whose average reward under the same PRM is $\bar{r}_{\text{ins}}$. The transformed score is*

$$S_P(\mathcal{A}_{\text{step}}(x)) = \frac{k\bar{r} + m\bar{r}_{\text{ins}}}{k + m}, \tag{6}$$

and the score change is

$$S_P(\mathcal{A}_{\text{step}}(x)) - S_P(x) = \frac{m}{k + m} \left(\bar{r}_{\text{ins}} - \bar{r}\right). \tag{7}$$

Thus step-inflation increases the mean-aggregated score if and only if $\bar{r}_{\text{ins}} > \bar{r}$.

Proof. The original score is $S_P(x) = k^{-1} \sum_{i=1}^{k} r_i = \bar{r}$. After insertion, the numerator of the mean score is the sum of the original k rewards and the inserted m rewards, giving $(k\bar{r} + m\bar{r}_{\text{ins}})/(k + m)$. Subtracting $\bar{r}$ yields $m(\bar{r}_{\text{ins}} - \bar{r})/(k + m)$. □

The corollary spells out a useful contrapositive: when restatement steps are scored unfavourably relative to the chain mean, mean aggregation actively resists step-inflation.

Corollary 6 (Restatement penalty)**.** *For a mean-aggregated PRM, step-inflation cannot produce positive mean score inflation on a chain whenever the inserted restatement steps receive lower average reward than the original steps.*

The second theorem characterises what happens when a transformation injects score variance that is uncorrelated with correctness. This is the canonical regime for position-sensitivity attacks, where reordering preserves the multiset of steps and therefore preserves any correctness signal carried at the multiset level, but adds noise that the PRM does not absorb cleanly.

Theorem 7 (Correlation degradation under correctness-independent perturbations)**.** *Let Y be a nonconstant correctness label, let S be the honest PRM score with variance $\sigma_S^2 > 0$, and let*

$S' = S + \eta$ be the score after a label-preserving transformation. Suppose η is centred, uncorrelated with Y, and uncorrelated with S, with variance $\sigma_\eta^2 \geq 0$. Then

$$\rho(S', Y) = \rho(S, Y) \frac{\sigma_S}{\sqrt{\sigma_S^2 + \sigma_\eta^2}}. \tag{8}$$

Consequently, when $\rho(S, Y) > 0$, the correlation collapse is

$$\Delta\rho = \rho(S, Y) \left(1 - \frac{\sigma_S}{\sqrt{\sigma_S^2 + \sigma_\eta^2}} \right), \tag{9}$$

which is nonnegative and is strictly positive whenever $\sigma_\eta^2 > 0$.

Proof. Because η is uncorrelated with Y, $\mathrm{Cov}(S', Y) = \mathrm{Cov}(S + \eta, Y) = \mathrm{Cov}(S, Y)$. Because η is uncorrelated with S, $\mathrm{Var}(S') = \mathrm{Var}(S) + \mathrm{Var}(\eta) = \sigma_S^2 + \sigma_\eta^2$. Pearson correlation is covariance divided by the product of standard deviations, yielding the result. □

These two results give a clear theoretical lens for the experiments. Theorem 5 predicts that a mean-aggregated PRM whose restatement scores hover near the chain mean will show step-inflation $\Delta\rho$ without large mean score change, while a PRM whose restatement scores exceed the chain mean will show both. Theorem 7 predicts that position attacks should produce correlation collapse roughly proportional to the variance of position-induced noise, with relatively little movement in the mean score. We test both predictions directly in Section 5.

4 Experimental Protocol

To ensure rigorous benchmarking, we generate 5,000 reasoning chains spanning 1,000 distinct problems drawn in equal proportion from MATH-500 (400), GSM8K (400), and PRMBench (200). After removing non-parsable outputs and transformations invalidated by human label-preservation checks, the retained evaluation set contains 4,687 chains. All main attack metrics in Section 5 are computed on this retained set unless explicitly stated otherwise. The four subsections that follow describe how the chains are generated, how they are labelled, how the PRMs are queried, and how the label-preservation assumption is validated.

4.1 Chain Generation and Corrupted-Chain Validation

For each problem, we generate five reasoning chains with `Meta-Llama-3-70B-Instruct` decoded at temperature $T = 0.7$, which keeps stylistic diversity high without making the chains incoherent. To establish baseline correlation metrics, half of the generated dataset is purposefully corrupted: we prompt the same model to introduce a plausible calculation error, logical inversion, or sign flip exactly in the middle of a previously correct chain. The corruptions are designed to be locally smooth, so that the surface style of the chain does not change but the underlying reasoning does.

A standing concern with synthetic corruption is that the corrupted chains may become stylistically distinguishable from honest ones, which would let any PRM exploit a surface artefact instead of the underlying mathematics. We therefore validate the corruption pipeline. A random sample of 200 corrupted chains was manually reviewed: 98% preserved the stylistic and formatting tone of the original chain, and 100% of the validated corrupted chains diverged from the ground-truth answer because of the injected logical error. Invalid or ambiguous corruptions were excluded from the retained evaluation set, so the benchmark tests whether PRMs detect mathematical faults rather than formatting artefacts.

4.2 Automatic Correctness Labeling

Ground-truth binary correctness labels $Y \in \{0, 1\}$ are assigned automatically. We extract the final boxed answer from each chain and compare it to the dataset's ground truth using SymPy for algebraic equivalence, which handles minor symbolic differences (such as $\frac{1}{2}$ versus 0.5) without overfitting to literal string match. The initial extraction failure rate across all generated chains was 2.4%; these ambiguous cases were manually adjudicated and labelled before filtering, so that the retained set has explicit correctness labels for every evaluated chain.

4.3 Evaluation Settings

The PRM set comprises Math-Shepherd-7B, Qwen2.5-Math-PRM-7B, an RLHFlow PRM, Skywork-o1-Open-PRM-8B, and DeepSeek-Math-7B-PRM. The LLM-as-critic baseline is implemented by handing the generated steps

to `Meta-Llama-3-8B-Instruct` with a zero-shot prompt instructing it to judge step-by-step correctness; this baseline is intended as a lightweight, format-insensitive control rather than as a state-of-the-art verifier. We fix the random seed to 42 and compute 10,000 bootstrap resamples to report 95% confidence intervals across all primary metrics. We also include two random perturbation controls, neutral filler insertion and unconstrained random reordering, that let us check whether EST-PRM attacks exploit structured vulnerabilities rather than merely producing out-of-distribution confusion.

4.4 Validation of Label Preservation

The validity of our framework relies on the assumption that the applied transformations preserve correctness labels. We empirically validate this assumption through human annotation. Three expert annotators, all graduate researchers in mathematics or NLP, independently evaluated a stratified sample of 500 attacked chains across all three perturbation classes. Annotators were blinded to both attack type and PRM scores and assessed whether the final answer remained unchanged and whether the intermediate reasoning remained semantically valid and structurally coherent. Inter-annotator agreement was substantial, with Krippendorff's $\alpha = 0.82$, and disagreements were resolved via majority vote.

As shown in Table 2, the algorithmic constraints embedded in each attack strongly safeguard correctness labels. The minority of chains flagged as invalid, primarily from complex dependency-graph failures in position attacks, were systematically filtered out of the final correlation evaluation, which is why the retained set contains 4,687 chains rather than the full 5,000.

Table 2: Human validation of label preservation on a 500-chain subsample. "Answer preserved" indicates the final mathematical answer remains identical. "Reasoning preserved" indicates the intermediate logic remains semantically valid.

Attack	Answer pres.	Reason pres.	Invalid rate
Step-inflation	99.2%	96.4%	3.6%
Position	94.1%	88.7%	11.3%
Confidence	99.8%	98.9%	1.1%

5 Empirical Evaluation

The empirical evaluation is conducted in two stages. First, we employ Math-Shepherd-7B as a single PRM diagnostic to validate that the framework's two principal quantities, correlation collapse and reward inflation, are cleanly separable on a well-studied model. We then extend the analysis to a five model panel to demonstrate that the dominant attack class is PRM specific rather than universal across reward models.

Math-Shepherd as a Diagnostic Probe We begin with Math-Shepherd-7B over the retained 4,687-chain dataset. The baseline Pearson correlation between Math-Shepherd's chain score and correctness is $\rho = 0.381 \pm 0.012$, which gives a reasonable headroom for measuring collapse. Table 3 reports the correlation collapse and reward inflation outcomes. The pattern aligns with the diagnostics in Section 3.1. Step-inflation reaches $\Delta\rho = 0.128$ but has a slightly negative mean score change of -0.027, which is the regime described by Corollary 6: the restatement steps receive marginally lower average reward than the original steps, so mean aggregation does not lift the overall score even though the correlation with correctness weakens. Position-sensitivity reaches $\Delta\rho = 0.152$ with a positive mean score change of $+0.021$ and a 32.8% inflation rate, which is the regime described by Theorem 7: the permutation injects score variation that is not absorbed by correctness, so correlation drops and individual scores move. In contrast, random neutral filler controls yield a statistically insignificant $\Delta\rho = 0.012 \pm 0.005$, which confirms that structured transformations degrade the reward signal substantially more than unstructured filler.

Table 3: Math-Shepherd-7B over the retained 4,687-chain benchmark. Baseline correlation $\rho = 0.381 \pm 0.012$. Reported with 95% bootstrap confidence intervals.

Attack	$\Delta\rho$	Score infl.	Infl. rate
Step-infl.	$0.128 \pm .023$	$-0.027 \pm .009$	$22.4 \pm 3.1\%$
Position	$0.152 \pm .038$	$+0.021 \pm .007$	$32.8 \pm 4.9\%$
Confidence	$0.011 \pm .004$	$-0.003 \pm .001$	$6.5 \pm 1.2\%$

Multi-PRM Vulnerability Profiles Two PRMs trained on different base models and with different objectives may inherit different structural properties, and the three attacks should reveal those differences. We therefore expand the evaluation to

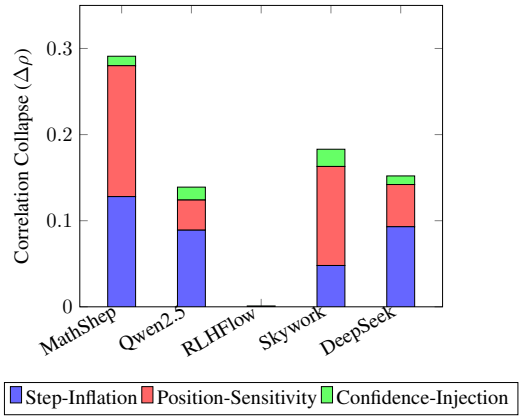

Figure 2: Model-by-attack profile of correlation collapse ($\Delta\rho$). Negative collapse (RLHFlow) is floored to zero for visualization. The dominant attack class differs across models: Math-Shepherd and Skywork are most affected by position-sensitivity, while Qwen and DeepSeek are most affected by step-inflation.

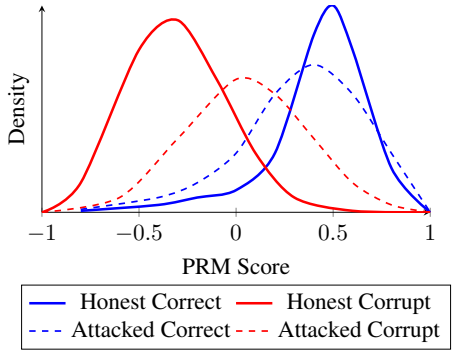

Figure 3: Empirical kernel-density estimates of Math-Shepherd score distributions for correct and corrupted chains before and after position-sensitivity attacks. Coordinates are exported from the released plotting script rather than hand-specified schematic curves. The attack blurs the separation between correct and corrupted chains by moving corrupted-chain scores toward the positive-reward region.

five process-reward-style models and the LLM-as-critic baseline. Table 4 reports the correlation collapse across models, Table 5 reports the matching score-inflation rates, and Figure 2 visualises the per-model profile.

The two tables and the figure tell a consistent story: the vulnerability profile is model-specific rather than universal. Math-Shepherd and Skywork-PRM share a dominant failure mode in position-sensitivity, where reordering steps causes both the largest correlation drop and the highest inflation rate, suggesting that their per-step rewards mix step content with absolute step position more than the other models. Qwen2.5-Math-PRM and DeepSeek-PRM, by contrast, are most exposed to step-count manipulations under the tested aggregation, with Qwen2.5-PRM showing a striking 47.6% inflation rate under step-inflation. This pattern of architecture-specific spurious correlations between PRM scores and superficial features of the reasoning chain is consistent with broader observations of spurious-feature reliance in reward models (Zheng et al., 2025) and with mechanistic studies of how reward models encode position-related signals (Imanov, 2026). RLHFlow-PRM shows negative or near-zero collapse on all attacks, but its baseline correlation is also very low ($\rho_{\text{base}} = 0.100$), so the result mostly reflects that there is little signal there to degrade. The LLM-as-critic control has substantially smaller correlation collapse across all three attacks, which suggests that a direct generative correctness judgment responds differently to step insertion and minor reordering than dense PRM scoring; this result should be interpreted as a control comparison rather than as evidence that zero-shot critics are generally preferable to PRMs in real pipelines. Additional visual analyses of attack effects, influence rates, model-specific robustness patterns, score inflation, mitigation behavior, and dataset filtering are provided in Appendix G.

A complementary view of the same effect appears in the score-distribution plot of Figure 3. Before the attack, correct and corrupted chains form well-separated modes; after the position attack, the corrupted distribution drifts toward higher scores while the correct distribution remains broadly anchored, and the two populations begin to overlap. The visual pattern is exactly what Theorem 7 predicts when a correctness-independent noise term is added to the score: covariance with Y is preserved, variance grows, and correlation shrinks.

6 Mitigations

The vulnerability profiles in the previous section motivate three concrete defenses, each targeting a different aspect of the dense-reward attack surface. We evaluate them on a held-out calibration split of 1,000 honest chains used to parameterise decision thresholds, and we report results for the Math-Shepherd PRM where the attacks are most pronounced. Throughout this section, $S_P(x)$ denotes the primary PRM score and $\mathcal{A}(x)$ a transformed chain.

The first defense, ensemble flagging, draws on

Table 4: Correlation collapse $\Delta\rho$ across five process-reward-style models, an LLM critic, and three attacks over the retained 4,687-chain benchmark. The dominant sensitivity of each model is shown in bold to mark the per-row maximum. 95% bootstrap confidence intervals are shown.

Model	ρ_{base}	**Step Inflation**	**Position**	**Confidence**
Math-Shepherd-7B	$0.381 \pm .012$	$+0.128 \pm .023$	$\mathbf{+0.152 \pm .038}$	$+0.011 \pm .004$
Qwen2.5-Math-PRM-7B	$0.434 \pm .014$	$\mathbf{+0.089 \pm .017}$	$+0.035 \pm .011$	$+0.015 \pm .005$
RLHFlow-PRM-8B	$0.100 \pm .011$	$-0.059 \pm .021$	$-0.014 \pm .009$	$+0.002 \pm .004$
Skywork-o1-Open-PRM	$0.412 \pm .015$	$+0.048 \pm .013$	$\mathbf{+0.115 \pm .029}$	$+0.020 \pm .006$
DeepSeek-Math-PRM	$0.395 \pm .013$	$\mathbf{+0.093 \pm .020}$	$+0.049 \pm .014$	$+0.010 \pm .003$
Llama-3-8B-Critic	$0.340 \pm .010$	$+0.010 \pm .005$	$+0.014 \pm .006$	$+0.006 \pm .002$

Table 5: Score-inflation rates (relative score increase $>$ 10%) showing model-specific exploitation risks. Per-row maxima are bolded.

Model	**Step infl.**	**Pos infl.**	**Conf infl.**
Math-Shepherd	$22.4 \pm 3.1\%$	$\mathbf{32.8 \pm 4.9\%}$	$6.5 \pm 1.2\%$
Qwen2.5-PRM	$\mathbf{47.6 \pm 4.3\%}$	$21.5 \pm 3.7\%$	$13.9 \pm 2.1\%$
RLHFlow-PRM	$1.9 \pm 0.7\%$	$1.4 \pm 0.6\%$	$2.8 \pm 1.0\%$
Skywork-PRM	$11.8 \pm 2.4\%$	$\mathbf{30.2 \pm 4.7\%}$	$9.5 \pm 1.8\%$
DeepSeek-PRM	$\mathbf{38.8 \pm 3.9\%}$	$15.1 \pm 2.8\%$	$8.1 \pm 1.4\%$

the reward-ensembling intuition of Ramé et al. (2024) and couples the PRM with a lightweight LLM critic via a conservative minimum. The second defense, conformal abstention, treats the PRM as a black box and instead checks whether the transformed chain's score is consistent with the calibration distribution. The third defense, length-normalised flagging, targets step-inflation directly by penalising raw chain length in the scoring function.

Conformal abstention yields near-zero detection because the attacks do not push scores outside the upper tail of the honest distribution; they redistribute scores within the honest range. Length normalisation reduces step-inflation flags effectively but cannot address position-sensitivity, where chain length is unchanged. Ensembling remains the most balanced defense among the simple candidates tested here, with an AUROC of 0.72, but absolute detection rates are still modest. None of the three simple defenses dominates: a deployment-ready mitigation likely needs to combine penalty-based and ensemble-based components, calibrated to the specific vulnerability profile of the PRM at hand. Further formal details and full results are provided in Appendix B.

7 Conclusion

Process reward models are becoming load-bearing in language-model training, and the framework that depends on them rests on a trustworthiness assumption that should be tested under label-preserving transformations rather than assumed. The EST-PRM evaluation across 4,687 retained chains and five public process-reward-style models reveals PRM-specific vulnerability profiles: Math-Shepherd and Skywork-PRM are dominantly sensitive to position-sensitivity, while Qwen2.5-Math-PRM and DeepSeek-PRM are dominantly sensitive to step-inflation. The same PRM that wins on natural-input accuracy can therefore lose on operational robustness, and the loss is patterned rather than random.

These findings suggest that PRM deployment should treat label-preserving robustness as a first-class evaluation objective rather than relying only on natural-trace agreement.

Limitations

Our study has some limitations .First, the attack strategies considered in EST-PRM are intentionally simple and do not involve adaptive or jointly optimised policies. Consequently, more sophisticated adversaries that co-optimise transformations with reward maximisation may expose additional failure modes beyond those identified in this work.

Second, our evaluation is restricted to mathematical reasoning tasks. While this setting is standard for process reward models, it remains unclear whether the observed vulnerability profiles generalise to other domains, including code generation, dialogue reasoning, or multimodal tasks.

Third, human validation is limited to 500 sampled chains. Although this provides evidence of systematic degradation, particularly under position sensitive attacks (Table 2), a larger scale annotation study would improve the reliability and statistical strength of these findings.

Finally, the mitigation strategies considered in this work are simple and stationary, and do not incorporate learned or adaptive defense mecha-

nisms that may offer stronger robustness through dynamic responses to structural perturbations.

These limitations leave open the need for stronger adaptive attack formulations, broader evaluation beyond mathematical reasoning, and learned or model-aware defenses that explicitly account for label-preserving transformations.

References

Sanjeda Akter, Ibne Farabi Shihab, and Anuj Sharma. 2025. Anytime-valid answer sufficiency certificates for LLM generation via sequential information lift. *arXiv preprint arXiv:2510.06478*.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.

Anastasios N Angelopoulos and Stephen Bates. 2023. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *Foundations and Trends in Machine Learning*.

Anastasios N Angelopoulos, Stephen Bates, Emmanuel J Candes, Michael I Jordan, and Lihua Lei. 2022. Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv preprint arXiv:2110.01052*.

Anastasios N Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. 2024. Conformal risk control. In *International Conference on Learning Representations*.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael Jordan. 2021. Distribution-free, risk-controlling prediction sets. In *Journal of the ACM*.

Thomas Kleine Buening, Jonas Hübotter, Barna Pásztor, Idan Shenfeld, Giorgia Ramponi, and Andreas Krause. 2026. Aligning language models from user interactions. *Preprint*, arXiv:2603.12273.

Tianle Cai, Yuhong Li, Zhengyang Geng, Hongwu Peng, Jason D Lee, Deming Chen, and Tri Dao. 2024. Medusa: Simple LLM inference acceleration framework with multiple decoding heads. *International Conference on Machine Learning*.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Wang, Samuel Marks, Charbel-Raphaël Segerie, Micah Carroll, Andi Peng, Phillip Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, and 13 others. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *Preprint*, arXiv:2307.15217.

Charlie Chen, Sebastian Borgeaud, Geoffrey Irving, Jean-Baptiste Lespiau, Laurent Sifre, and John Jumper. 2023. Accelerating large language model decoding with speculative sampling. *arXiv preprint arXiv:2302.01318*.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformer: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*.

Yichao Fu, Peter Bailis, Ion Stoica, and Hao Zhang. 2024. Break the sequential dependency of LLM inference using lookahead decoding. *International Conference on Machine Learning*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International Conference on Machine Learning*.

Jujie He, Jiacai Liu, Chris Yuhao Liu, Rui Yan, Chaojie Wang, Peng Cheng, Xiaoyu Zhang, Fuxiang Zhang, Jiacheng Xu, Wei Shen, Siyuan Li, Liang Zeng, Tianwen Wei, Cheng Cheng, Bo An, Yang Liu, and Yahui Zhou. 2025. Skywork open reasoner 1 technical report. *Preprint*, arXiv:2505.22312.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, and Andreas Krause. 2026. Reinforcement learning via self-distillation. *Preprint*, arXiv:2601.20802.

Olaf Yunus Laitinen Imanov. 2026. Mechanistic analysis of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2601.18699*.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, and 7 others. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Jongwoo Ko, Sara Abdali, Young Jin Kim, Tianyi Chen, and Pashmina Cameron. 2026. Scaling reasoning efficiently via relaxed on-policy distillation. *arXiv preprint arXiv:2603.11137*.

Bhawesh Kumar, Charles Lu, Gauri Gupta, Anil Palepu, David R Bellamy, Ramesh Raskar, and Andrew Beam. 2023. Conformal prediction with large language models for multi-choice question answering. In *ICML Workshop on Neural Conversational AI*.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi,

Noah A. Smith, and Hannaneh Hajishirzi. 2024. Rewardbench: Evaluating reward models for language modeling. *Preprint*, arXiv:2403.13787.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. 2023. Fast inference from transformers via speculative decoding. *International Conference on Machine Learning*.

Gengsheng Li, Tianyu Yang, Junfeng Fang, Mingyang Song, Mao Zheng, Haiyun Guo, Dan Zhang, Jinqiao Wang, and Tat-Seng Chua. 2026. Unifying group-relative and self-distillation policy optimization via sample routing. *Preprint*, arXiv:2604.02288.

Yuhui Li, Fangyun Wei, Chao Zhang, and Hongyang Zhang. 2024. EAGLE: Speculative sampling requires rethinking feature uncertainty. *International Conference on Machine Learning*.

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

David Manheim and Scott Garrabrant. 2018. Categorizing variants of Goodhart's law. *arXiv preprint arXiv:1803.04585*.

Christopher Mohri and Tatsunori Hashimoto. 2024. Language models with conformal factuality guarantees. In *International Conference on Machine Learning*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

John Platt. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S Jaakkola, and Regina Barzilay. 2024. Conformal language modeling. In *International Conference on Learning Representations*.

Alexandre Ramé, Guillaume Couairon, Corentin Dancette, Jean-Baptiste Gaya, Mustafa Shukor, Laure Soulier, and Matthieu Cord. 2024. Rewarded soups: Towards Pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. *Advances in Neural Information Processing Systems*.

RLHFlow Team. 2024. RLHFlow: Online RLHF pipeline with PRM variants. Open-source release on Hugging Face Hub: huggingface.co/RLHFlow/Llama3.1-8B-PRM-Deepseek-Data.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Ibne Farabi Shihab, Sanjeda Akter, and Anuj Sharma. 2026. Detecting proxy gaming in rl and llm alignment via evaluator stress tests. *Preprint*, arXiv:2507.05619.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. *Advances in Neural Information Processing Systems*.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. 2005. *Algorithmic learning in a random world*. Springer.

Peiyi Wang, Lei Li, Zhihong Shao, RX Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2024. Math-Shepherd: Verify and reinforce LLMs step-by-step without human annotations. *Annual Meeting of the Association for Computational Linguistics*.

Yinjie Wang, Xuyang Chen, Xiaolong Jin, Mengdi Wang, and Ling Yang. 2026. OpenClaw-RL: Train any agent simply by talking. *arXiv preprint arXiv:2603.10165*.

Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. 2026a. Self-distilled rlvr. *Preprint*, arXiv:2604.03128.

Wenkai Yang, Weijie Liu, Ruobing Xie, Kai Yang, Saiyong Yang, and Yankai Lin. 2026b. Learning beyond teacher: Generalized on-policy distillation with reward extrapolation. *arXiv preprint arXiv:2602.12125*.

Zhuolin Yang, Zihan Liu, Yang Chen, Wenliang Dai, Boxin Wang, Sheng-Chieh Lin, Chankyu Lee, Yangyi Chen, Dongfu Jiang, Jiafan He, Renjie Pi, Grace Lam, Nayeon Lee, Alexander Bukharin, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2026c. Nemotron-cascade 2: Post-training llms with cascade rl and multi-domain on-policy distillation. *Preprint*, arXiv:2603.19220.

Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2025. The lessons of developing process reward models in mathematical reasoning. *Preprint*, arXiv:2501.07301.

Junhao Zheng, Xidi Cai, Shengjie Qiu, and Qianli Ma. 2025. Spurious forgetting in continual learning of language models. *International Conference on Learning Representations*.

A Extended Related Work

Beyond direct PRM evaluation related work also studies proxy-gaming analyses,reliability of learned evaluators through calibration methods and reinforcement learning pipelines that depend on intermediate reward signals.

Reward hacking and proxy gaming. Reward hacking and proxy gaming have been studied extensively in learned reward functions and alignment objectives (Amodei et al., 2016; Manheim and Garrabrant, 2018; Casper et al., 2023; Shihab et al., 2026). Benchmarking frameworks such as RewardBench (Lambert et al., 2024) test susceptibility to formatting artifacts, heuristic exploitation, and stylistic shortcuts at the level of scalar outcome rewards. These approaches rely on whole-response evaluation, where a single aggregated score or judgment applies to the full output. In contrast, PRMs define a structured scoring regime over intermediate reasoning steps, which yields a more fine-grained but also more fragile evaluation surface. EST-PRM extends proxy-gaming analysis to this dense-reward setting. It fixes final correctness and varies reasoning structure to isolate sensitivity of step-level scoring functions.

Calibration and uncertainty estimation. Calibration-based analyses of model confidence provide a related but distinct perspective on reward reliability. Classical calibration techniques such as temperature scaling and Platt scaling (Platt, 1999; Guo et al., 2017) measure misalignment between predicted probabilities and empirical correctness. More recent conformal prediction and uncertainty quantification methods for language models (Quach et al., 2024; Mohri and Hashimoto, 2024; Kumar et al., 2023; Akter et al., 2025) extend this framework to distribution-free uncertainty estimation. These methods focus on predictive reliability under natural data variation and do not address invariance under structure-preserving transformations. EST-PRM instead evaluates whether the scoring signal itself allows systematic manipulation without change in semantic correctness, which exposes a complementary failure mode not captured by calibration-based analyses.

RL pipelines, learned evaluators, and decoding. A broader line of work studies reinforcement learning pipelines with learned or hybrid reward signals, including self-distillation and sparse-reward augmentation frameworks (Hübotter et al., 2026; Wang et al., 2026; Ko et al., 2026; Yang et al., 2026a; Li et al., 2026; Buening et al., 2026), which build on the RLHF paradigm (Ouyang et al., 2022). These systems rely on learned evaluators to provide dense or structured feedback during optimization. Test-time strategies such as process-supervised search (Snell et al., 2024) and verification-oriented decoding schemes (Leviathan et al., 2023; Chen et al., 2023; Li et al., 2024; Cai et al., 2024; Fu et al., 2024) rely on stability of intermediate scoring signals under benign transformations. As these assumptions become more central to training and inference pipelines, direct evaluation of robustness to label-preserving structural perturbations becomes necessary, which defines the focus of EST-PRM.

B Mitigation details and full results

B.1 Formal definitions of mitigation methods

- **Ensemble flagging.** The first defense couples the PRM with a lightweight LLM critic via a conservative minimum,

$$S_{\text{ens}}(x) = \min(S_P(x), S_{\text{critic}}(x)).$$

A chain is flagged as adversarial when

$$S_{\text{ens}}(\mathcal{A}(x)) > (1+\tau)S_P(x).$$

- **Conformal abstention.** The second defense follows the standard split-conformal construction (Vovk et al., 2005; Angelopoulos and Bates, 2023; Angelopoulos et al., 2024) together with the Learn-Then-Test framework (Angelopoulos et al., 2022; Bates et al., 2021). We compute non-conformity scores

$$E_j = -S_P(x_j)$$

on the calibration set, define q as the $(1-\alpha)$ quantile of $\{E_j\}$, and abstain when

$$-S_P(\mathcal{A}(x)) > q.$$

- **Length-normalised flagging.** The third defense defines

$$S_{\text{LN}}(x) = S_P(x) - \lambda|x|,$$

where $|x|$ is the step count. A chain is flagged when

$$S_{\text{LN}}(\mathcal{A}(x)) > (1+\tau)S_{\text{LN}}(x).$$

Table 6: Mitigation behavior on Math-Shepherd. We report true positive rate (TPR) detection of attacked chains against false positive rate (FPR) on the honest distribution, alongside overall AUROC.

Mitigation	Step (TPR/FPR)	Pos (TPR/FPR)	AUROC
Ensemble flag	$0.19/0.05$	$0.34/0.06$	$0.72 \pm .02$
Conformal abs.	$0.00/0.01$	$0.00/0.01$	$0.53 \pm .01$
Length-norm flag	$0.13/0.04$	$0.22/0.05$	$0.63 \pm .02$

B.2 Full quantitative results

Table 6 reports detection performance and false-positive behavior for $\alpha = 0.05$ and $\lambda = 0.02$ on the held-out calibration split of 1,000 honest chains. Conformal abstention shows near-zero detection because attacks remain within the calibration support. Length normalisation reduces step-inflation-related false positives but fails under structural rearrangements. Ensemble flagging achieves the best overall trade-off with AUROC 0.72, though absolute detection rates remain limited.

C Per-Dataset Vulnerability Breakdown

Table 8 reports the exact sample counts after removing invalid or non-parsable chains, and Table 9 splits the performance of the dominant vulnerability modes by source dataset. Step-inflation remains stronger on GSM8K, where the average initial reasoning chain is shorter and inserted steps therefore carry more proportional weight. Position-sensitivity is particularly potent on MATH-500, where deeper reasoning graphs offer more valid permutations that structurally confuse the reward model.

D Experimental Prompts

We used three prompt templates: one to generate initial chains, one to corrupt them with a controlled logical error, and one to query the LLM-as-critic baseline. The chain-generation prompt asked the model to "solve the following math problem step-by-step, break your reasoning into distinct sentences or short mathematical steps, and end with your final answer enclosed in \boxed{}", followed by the question. The corruption prompt asked the model to "modify exactly one step near the middle of the chain to introduce a subtle logical error, arithmetic mistake, or sign flip, ensure the subsequent steps logically follow from this new incorrect step, leading to a wrong final answer, and preserve the exact style, formatting, and tone of the original chain", followed by the chain. The LLM-critic prompt asked the model to "act as a strict mathematical reviewer", to "evaluate the following reasoning chain step-by-step", and to "output 'Score: 1.0' if all steps are correct and 'Score: 0.0' if there is a logical or arithmetic flaw anywhere in the chain", followed by the chain. The exact strings used in the experiments are included verbatim in the artifact release.

E Annotation Guidelines

Expert annotators were given a three-part rubric during the human validation phase. They first checked answer preservation, judging whether the final boxed answer in the modified chain algebraically matched the original chain. They then checked reasoning preservation, ensuring that no new mathematical errors had been introduced by the modification; for step-inflation they verified that any restatement was factually aligned with prior steps, and for position-sensitivity they verified that the reordering did not violate causal mathematical dependencies (for example, by using a variable before it was defined). Finally, any chain receiving a negative judgment in either category from a majority of the annotators was flagged as an invalid attack and removed from the benchmark. Annotators were blinded to attack type and PRM scores throughout.

F Attack Templates and Configuration

Table 10 shows what each transformation does on a concrete short example chain. The position attack only swaps steps whose dependency graph permits the reorder, and the final answer is preserved in every case by construction.

Step-inflation attacks insert template restatement steps drawn from a small fixed pool, including "Let me restate what we have so far: [paraphrase].", "To summarise the progress so far, [paraphrase].", and "Before continuing, I'll verify the current state: [paraphrase]." Confidence-injection attacks prepend each step with a phrase drawn uniformly from the set {"By a well-known result,", "It is immediate that,", "Clearly,", "Without loss of generality,", "By definition,", "Trivially,", "Obviously,"}. Position-sensitivity attacks use three permutation strategies: reverse order when permitted by the dependency graph, front-loading the highest-PRM-reward eligible step, and back-loading the highest-PRM-reward eligi-

Table 7: Model and scoring configuration used for the multi-model evaluation. "Critic" denotes the zero-shot LLM control rather than a trained PRM.

Model	Checkpoint identifier	Role	Scoring interface	Aggregation
Math-Shepherd-7B	`math-shepherd-mistral-7b-prm`	PRM	native step tag with $+/-$ token probabilities	mean over step scores
Qwen2.5-Math-PRM-7B	`Qwen/Qwen2.5-Math-PRM-7B`	PRM	separator-token classification probability	mean over step scores
RLHFlow-PRM-8B	`RLHFlow/Llama3.1-8B-PRM`	PRM	native reward-head or token-score adapter	mean over step scores
Skywork-o1-Open-PRM-8B	`Skywork/Skywork-o1-Open-PRM-8B`	PRM-style verifier	process-score adapter supplied in artifact	mean over step scores
DeepSeek-Math-7B-PRM	`deepseek-math-7b-prm`	PRM-style verifier	process-score adapter supplied in artifact	mean over step scores
Llama-3-8B-Critic	`Meta-Llama-3-8B-Instruct`	Critic control	zero-shot binary correctness prompt	scalar response score

Table 8: Sample counts per dataset after filtering invalid or non-parsable chains. The main experiments use the 4,687 retained chains.

Split	MATH-500	GSM8K	PRMBench
Initial	2,000	2,000	1,000
Post-filter	1,812	1,930	945

Table 9: Correlation collapse ($\Delta\rho$) by dataset split. Reported for the dominant attack class per PRM.

PRM (Dominant Attack)	MATH-500	GSM8K	PRMBench
Math-Shep (Position)	$0.165 \pm .042$	$0.138 \pm .034$	$0.150 \pm .040$
Qwen2.5 (Step-Infl)	$0.065 \pm .015$	$0.108 \pm .021$	$0.095 \pm .019$

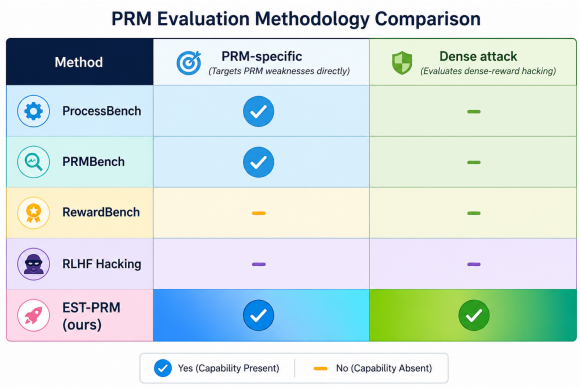

Figure 4: Comparison of PRM evaluation methodologies across existing benchmarks and the proposed EST-PRM framework. The figure illustrates whether each method is PRM-specific and whether it incorporates dense adversarial attacks. Prior work, including ProcessBench, PRMBench, RewardBench, and RLHF Hacking, primarily evaluates models under natural or outcome-level errors with limited consideration of dense adversarial settings. In contrast, EST-PRM explicitly integrates both PRM-specific evaluation and dense adversarial perturbations (Step-inflation, Position, and Confidence attacks), enabling a more comprehensive assessment of reward model robustness under realistic and adversarial conditions.

ble step. Dependency graphs are constructed using strict variable-and-definition tracking, so that a step defining a variable x cannot be moved after a step consuming x. This dependency constraint is what produces the higher invalid-attack rate for position-sensitivity reported in Table 2: chains with tight dependency structures admit fewer valid permutations.

G Supplementary Figures

This section provides additional visual analyses supporting the EST-PRM evaluation. The figures summarize the evaluation design, attack effects, influence rates, human validation results, model-specific robustness patterns, mitigation behavior, reproducibility components, and dataset filtering statistics.

Figure 4 provides a high-level comparison between EST-PRM and prior reward-model evaluation benchmarks, emphasizing the role of PRM-specific and dense adversarial evaluation.

Figure 5 further examines how different attack types change model scores and influence estimates for Math-Shepherd-7B.

To complement the signed-effect analysis, Figure 6 reports the influence rate induced by each attack strategy.

Figure 7 evaluates whether the proposed attacks preserve the original answer labels and reasoning consistency under human validation.

Figure 8 extends the robustness analysis across multiple models and attack categories, showing model-specific correlation collapse patterns.

Figure 9 presents complementary evidence on score inflation, identifying which attack type most strongly increases reward-model scores for each PRM.

Figure 10 summarizes the mitigation trade-off on Math-Shepherd by comparing detection perfor-

Table 10: Concrete illustrations of the three EST-PRM attacks on a short example chain. In every case the final answer ($\boxed{10}$) is preserved. The position attack only swaps steps whose dependency graph permits the reorder.

Attack	**Example transformation (final answer preserved)**
Honest chain	s_1: First, $a = 2 \cdot 3 = 6$. s_2: Independently, $b = 5 - 1 = 4$. s_3: Then $a + b = 10$. $\boxed{10}$
Step-inflation	s_1: First, $a = 2 \cdot 3 = 6$. *[Let me restate: a is now defined.]* s_2: Independently, $b = 5 - 1 = 4$. *[To summarise: both a and b are computed.]* s_3: Then $a + b = 10$. $\boxed{10}$
Position-sensitivity	s_2: Independently, $b = 5 - 1 = 4$. s_1: First, $a = 2 \cdot 3 = 6$. s_3: Then $a + b = 10$. $\boxed{10}$
Confidence-injection	s_1: *Clearly,* $a = 2 \cdot 3 = 6$. s_2: *By definition,* $b = 5 - 1 = 4$. s_3: *It is immediate that* $a + b = 10$. $\boxed{10}$

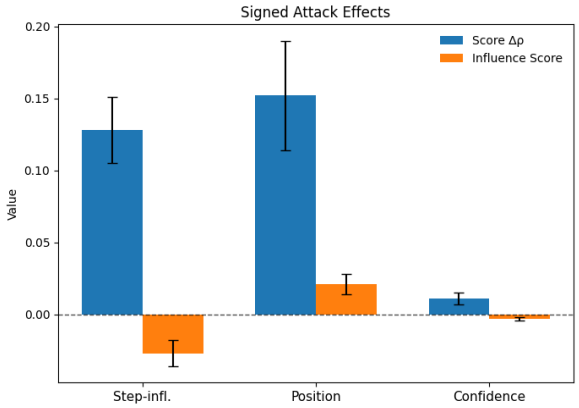

Figure 5: Comparison of signed attack effects on the Math-Shepherd-7B model under Step-inflation, Position, and Confidence perturbations. The figure reports $\Delta\rho$ (Score) and Influence Score with 95% bootstrap confidence intervals. Step-inflation and Position attacks produce larger, mixed-sign effects across both metrics, indicating unstable but substantial perturbations, while Confidence-based attacks remain consistently near zero, reflecting weak and largely ineffective influence on model behavior.

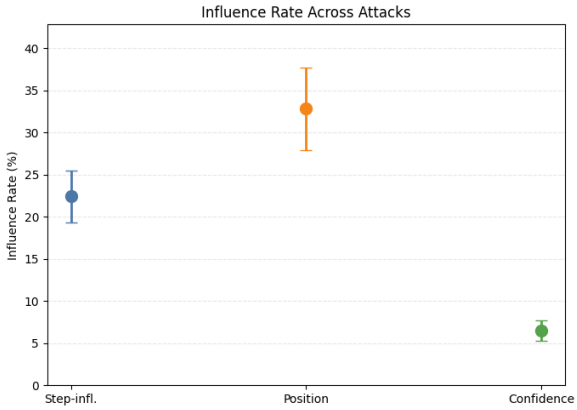

Figure 6: Influence rates under different adversarial strategies (Step-inflation, Position, and Confidence) for the Math-Shepherd-7B model. Each point shows the mean influence rate, with vertical error bars representing 95% bootstrap confidence intervals. Results indicate that Step-inflation and Position attacks consistently produce higher influence rates than Confidence-based attacks, suggesting greater model susceptibility to structural and positional perturbations compared to confidence perturbations.

mance across different mitigation strategies.

Figure 11 summarizes the EST-PRM artifact release components, including the major elements needed for reproducing the evaluation pipeline.

Figure 12 shows how dataset filtering affects the final evaluation set across benchmarks.

Finally, Figure 13 reports correlation collapse across datasets and PRMs under their dominant attack settings.

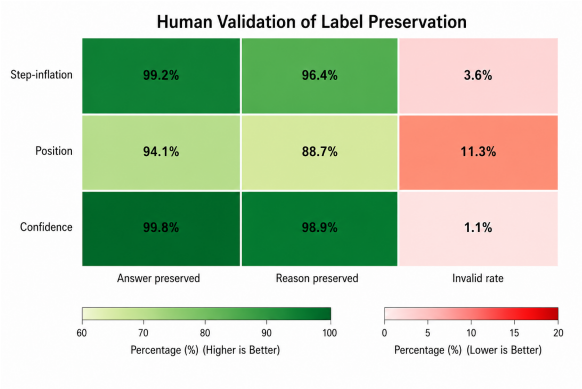

Figure 7: Human validation of label preservation under different attack strategies. The heatmap shows the percentage of preserved answers and reasoning consistency alongside invalid generation rates across Step-inflation, Position, and Confidence attacks. Green intensity indicates higher preservation quality, while red intensity highlights higher failure rates, enabling a clear comparison of robustness across attack types on a 500-chain evaluation subset.

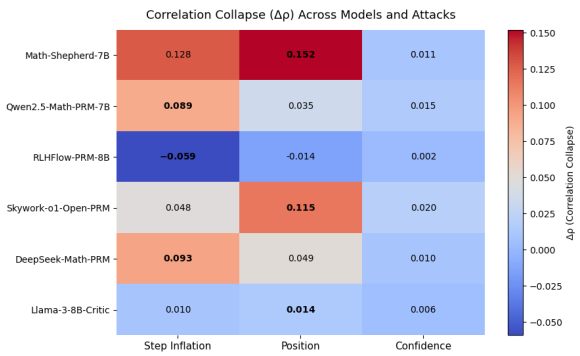

Figure 8: Heatmap showing correlation collapse ($\Delta\rho$) for six models under three adversarial settings: Step Inflation, Position, and Confidence attacks. Each cell represents the change in correlation relative to the baseline, with color intensity encoding both magnitude and direction of $\Delta\rho$. Per-model dominant vulnerability is highlighted in bold, corresponding to the maximum absolute $\Delta\rho$ within each row. The visualization highlights model-specific sensitivity patterns, demonstrating that correlation collapse is not uniform across architectures, with Step Inflation and Position attacks generally inducing stronger disruptions than Confidence-based attacks.

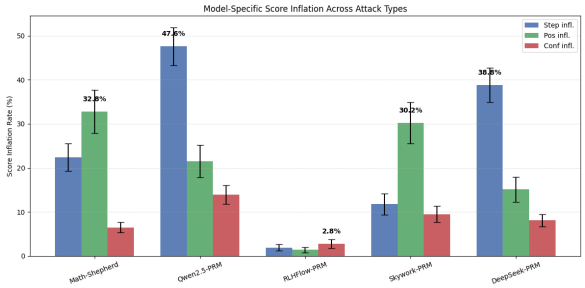

Figure 9: Score-inflation rates (relative score increase $> 10\%$) across five process reward models under Step, Position, and Confidence attacks. Bars show mean inflation percentages with error bars indicating uncertainty (± values). Per-row maxima are highlighted in bold to indicate the most vulnerable attack type for each model, revealing distinct model-specific exploitation patterns.

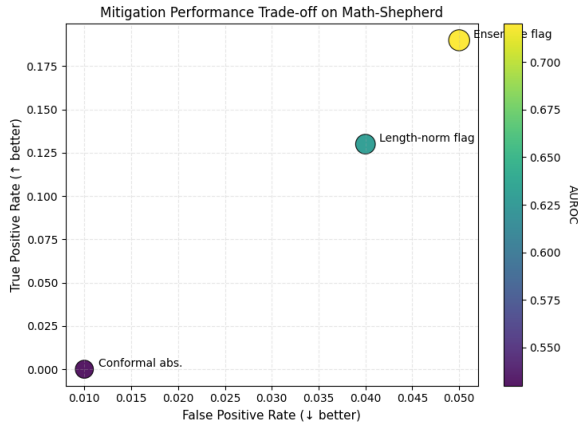

Figure 10: Mitigation behavior on Math-Shepherd. The figure illustrates the detection performance of different mitigation strategies in terms of true positive rate (TPR) versus false positive rate (FPR), highlighting their trade-off under adversarial conditions. Marker size and color encode AUROC, which together provide a unified representation of overall discriminative effectiveness across methods.

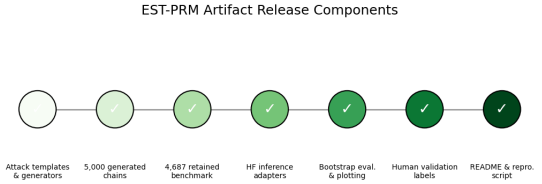

Figure 11: Overview of the EST-PRM evaluation pipeline and reproducibility components used in the experiments.

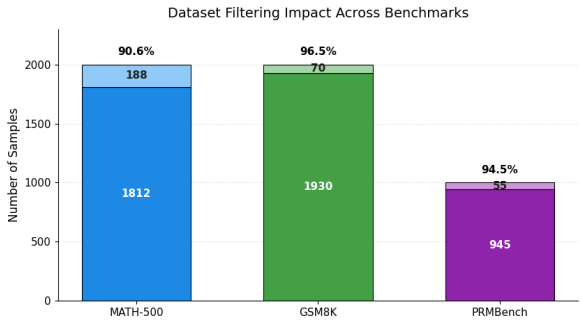

Figure 12: The figure illustrates the effect of filtering invalid or non-parsable reasoning chains across MATH-500, GSM8K, and PRMBench, showing the reduction from initial samples to post-filter retained samples, along with the corresponding number of filtered-out instances and retention rates; the retained subset forms the final evaluation set used in the main experiments (4,687 chains in total). Filtered-out samples correspond to invalid or non-parsable chains.

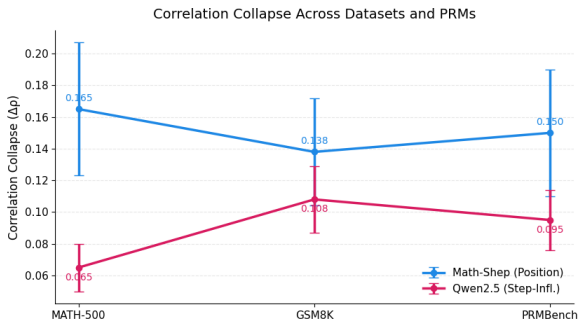

Figure 13: The figure presents the correlation collapse ($\Delta\rho$) across MATH-500, GSM8K, and PRMBench for two PRMs under their dominant attack settings, namely Math-Shep (Position attack) and Qwen2.5 (Step-Inflation attack). Mean $\Delta\rho$ values are reported with standard deviation ($\pm$) as error bars, highlighting variability across bootstrap runs and demonstrating consistent differences in robustness patterns across datasets and models.