SCIPRI: PROCESS-REWARD INVERSION IN SCIENTIFIC AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Process reward models (PRMs) score the intermediate steps of language-model agents and increasingly drive step selection, tree search, and reinforcement learning. Almost all public PRMs are trained on *mathematical* reasoning, where a step's value is its local arithmetic/logical correctness. Scientific agents (data analysis, experiment design, ML engineering) instead produce steps whose value is the *information* they contribute toward a correct scientific conclusion, often carried by terse, information-dense actions (code, observations, exploratory tests) rather than fluent prose. We introduce **SciPRI**, a measurement-and-mitigation framework that quantifies per-step information gain VIG_t with a calibrated probe over *frozen* hidden states (the change in the calibrated log-probability of task success) and uses it to audit reward models. Our central finding is **process-reward inversion**: the most widely used PRMs, math-trained *discriminative* reward models, become *negatively* aligned with information gain on scientific steps. The same model flips sign between domains: Qwen2.5-Math-PRM has length-controlled partial correlation $\rho{=}{+}0.209$ and mixed-effects slope $\beta_I{=}{+}0.0094$ ($p{<}10^{-29}$) between reward and information gain on a math control, but $\rho{=}{-}0.170$, $\beta_I{=}{-}0.0101$ ($p{=}3\times10^{-42}$) on scientific tasks. The inversion is gold-validated on 102 ScienceAgentBench tasks (math PRMs correlate -0.132 to $+0.045$ with gold scientific step quality, where they should be strongly positive), is robust across six generator models, three alternative observer models, and many estimator choices, and harms Best-of-N selection below random. A panel of reward models reveals a spectrum of failure: discriminative math PRMs invert, value-head PRMs collapse, and a strong general LLM-critic mis-selects. We introduce **InfoPRM**, a lightweight reward head trained to track information gain and predict the outcome while remaining length-decorrelated; it matches the best Best-of-N result on BLADE (0.711) and wins outright on ScienceAgentBench (0.810 vs. 0.749), agrees with gold scientific step labels ($+0.294$), and, in reward-guided generation, improves a weak agent's task success by 16.1–18.9 percentage points. SciPRI argues that process supervision for scientific agents must be measured against information gain, not against resemblance to mathematical reasoning in actual scientific-agent practice.

1 INTRODUCTION

Process reward models (PRMs) assign a scalar quality $r_t \in [0, 1]$ to each step of a multi-step solution and have become central infrastructure for reasoning and agentic LLMs: they rerank candidates, guide beam/tree search, prune incorrect branches early, and shape dense rewards for reinforcement learning (Lightman et al., 2024; Wang et al., 2024; Zhang et al., 2025). Their supervision is overwhelmingly *mathematical*: PRM800K-style human labels and Math-Shepherd-style Monte-Carlo rollouts define a step's value as its local correctness within an arithmetic or symbolic derivation. A growing literature pushes this paradigm further (generative PRMs that reason about each step (Khalifa et al., 2025; Zhao et al., 2026; She et al., 2025), universal/ensemble PRMs (Tan et al., 2026), active-learning-efficient PRMs (Duan et al., 2025), and min-form credit assignment (Cheng et al., 2026)) almost entirely benchmarked on math (ProcessBench (Zheng et al., 2025a), PRMBench (Song et al., 2025)), leaving transfer to scientific-agent reasoning largely unresolved.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

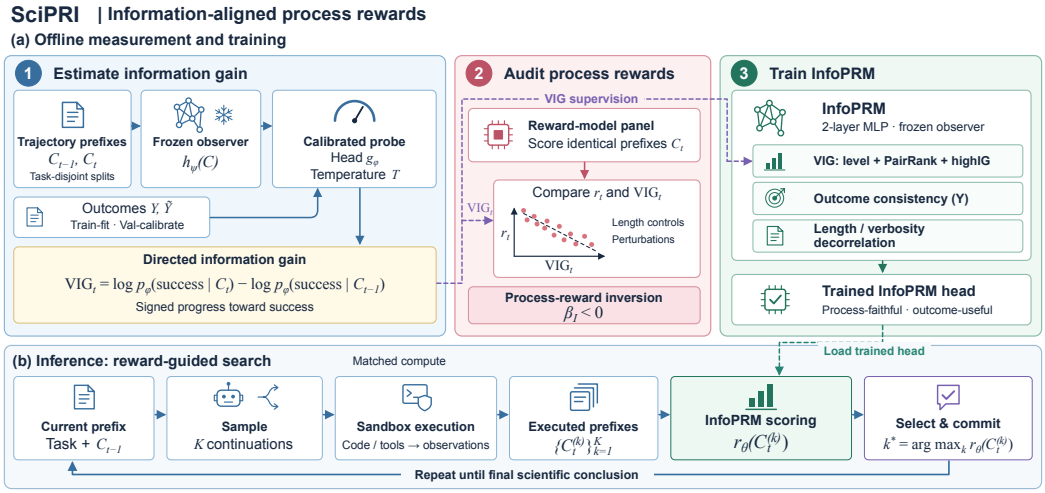

Figure 1: **Overview of SciPRI.** Offline, scientific-agent trajectory prefixes and task outcomes supervise a calibrated probe over frozen observer features; the resulting directed information gain VIG_t is compared with reward-model scores r_t to expose process-reward inversion and, together with outcome and length-decorrelation signals, train InfoPRM. At inference, the agent executes K candidate continuations in a sandbox, InfoPRM ranks their resulting prefixes, and the highest-reward step is committed to advance the trajectory toward its final scientific conclusion under matched compute.

Scientific agents value steps that *reduce uncertainty about the correct scientific conclusion*: validating data, fitting a statistical model, ruling out a hypothesis, or exposing a key result. Such information often arrives through terse actions: a two-line `statsmodels` call (Seabold et al., 2010), an error, or a printed coefficient. Prior work documents out-of-domain PRM degradation (Song et al., 2025; Li et al., 2026), verbosity bias, and proxy exploitation (Zheng et al., 2025b; Fein et al., 2026; Liu et al., 2025). What remains unmeasured is the per-step alignment of transferred rewards with task-relevant information on scientific trajectories, and its consequences for agent selection.

SciPRI (Fig. 1) measures this alignment using the change in task-success log-probability from a calibrated frozen-state probe, validated against gold step labels. We find **process-reward inversion**: the same PRM that rewards information gain on math *anti-rewards* it on BLADE. Complementary gold-label evidence spans 102 ScienceAgentBench tasks; the sign flip survives alternative generators, observers, and estimators. The failure also affects decisions: PRM selection falls below random in Best-of-N, and inverted guidance barely improves over random during matched-generation search on held-out BLADE tasks under the same candidate-generation and sandboxed execution protocol.

Contributions.

1. **A frozen-probe information-gain estimator**, calibrated and validated against gold step labels in mathematical and scientific reasoning under held-out task evaluation (Sec. 4.1).

2. **An alignment diagnostic suite**: length-controlled correlation, inversion rate, anti-information reward (AIR), high-information recall, and a mixed-effects slope, plus perturbations with explicit length accounting for each controlled scientific-step edit (Sec. 4.2).

3. **Evidence of process-reward inversion**: a math-to-BLADE sign flip across reward models and estimator choices, complemented by weak or negative agreement with gold scientific step labels on SAB and downstream candidate-selection failures on BLADE tasks (Sec. 6).

4. **InfoPRM**, a deployable reward head that ties for best on BLADE, wins on SAB Best-of-N, and improves agents during reward-guided search with matched compute (Sec. 4.4, 4.5).

2 Related Work

Process reward models. Step-level verifiers were popularized by PRM800K (Lightman et al., 2024) and Math-Shepherd (Wang et al., 2024), with strong open discriminative models from Qwen2.5-Math-PRM (Zhang et al., 2025), RLHFlow (Xiong et al., 2024), Skywork-o1 (He et al., 2024), and EurusPRM (Cui et al., 2025). Generative PRMs reason before scoring (ThinkPRM (Khalifa et al., 2025), GenPRM (Zhao et al., 2026), R-PRM (She et al., 2025)); universal and actively-trained PRMs push ProcessBench/ PRMBench SOTA (AURORA/Universal-PRM (Tan et al., 2026), ACTPRM (Duan et al., 2025)). ReasonEval (Xia et al., 2025) separates validity from redundancy. All target math; PRMBench (Song et al., 2025) broadens fine-grained error evaluation but remains centered on mathematical reasoning and scores PRM *accuracy*, not alignment with per-step information gain during scientific-agent execution on shared action and observation prefixes.

Agentic and scientific reward models. Recent work adapts PRMs to agents whose steps lack clear-cut correctness: early AgentPRM work uses Monte Carlo rollout targets (Choudhury, 2025), while subsequent work models step-wise promise and progress via TD/GAE (Xi et al., 2026). Web-Shepherd scores web-navigation trajectories (Chae et al., 2026), and ToolPRMBench benchmarks tool-use PRMs (Li et al., 2026). Closest to us, DataPRM (Qiu et al., 2026) observes that general-domain PRMs "fail to detect silent errors … and erroneously penalize exploratory actions" on data analysis, and trains a generative ternary PRM in response. SciPRI is complementary and goes further. We (i) *quantify* the failure as a sign-flipped, length-controlled *information-gain inversion* with a frozen-probe estimator and a diagnostic suite, gold-validated and robust across generators/observers; (ii) show it harms Best-of-N and reward-guided generation; and (iii) provide InfoPRM, an explicitly information-aligned, length-decorrelated head (we include a DataPRM-style ternary reward as a directly comparable baseline under matched evaluation on the shared features).

Reward hacking, length, and verbosity bias. Length/verbosity exploitation is a canonical failure of learned rewards (Zheng et al., 2025b; Fein et al., 2026; Liu et al., 2025), and reward hacking generalizes across domains (MacDiarmid et al., 2025). We show the scientific-step failure is *not* a verbosity bias (PRMs in fact penalize padding) but a representational domain mismatch in science.

Scientific-agent benchmarks. BLADE (Gu et al., 2024), ScienceAgentBench (Chen et al., 2025), DiscoveryWorld (Jansen et al., 2024), MLE-bench (Chan et al., 2025), and SciAgentGym (Shen et al., 2026) evaluate data-driven scientific agents; we use the first four, add gold step labels to the ScienceAgentBench trajectories, and use ProcessBench as a fixed mathematical reasoning control.

Information-theoretic credit assignment. Treating a step's value as information gain connects to Bayesian experimental design (Lindley, 1956) and value-of-information analysis (Howard, 1966). SciPRI operationalizes this for process supervision via a frozen-probe estimator of task progression.

3 Problem Setup

A *trajectory* $\tau = (\text{prompt}, s_1, \ldots, s_T)$ is one agent run on a task, where each step s_t has text, an optional action (code / tool call), an observation, and environment-state deltas. Each benchmark defines a scientific target Y (Sec. 5). A reward model assigns $r_t \in [0, 1]$ to each step. We write the serialized prefix, including the task prompt, all steps thus far, and the current environment state,

$$C_t \;=\; \text{serialize}(\text{prompt},\, s_{1:t},\, \text{env}_t)\,, \qquad C_0 = \text{prompt}, \tag{1}$$

and require that every reward model and our probe score *the same* prefixes, so that r_t and our information-gain score are directly comparable at every step by construction on the same agent task.

4 Method

SciPRI has four components (Fig. 1): (i) a frozen-probe estimator of per-step information gain VIG_t; (ii) a diagnostic suite that relates any reward model's r_t to VIG_t; (iii) a controlled perturbation test; and (iv) InfoPRM, a deployable information-aligned reward head for scientific agents, evaluated through offline candidate selection and online sandboxed search under matched compute.

4.1 INFORMATION GAIN FROM A FROZEN PROBE

Calibrated outcome probe. Let $h_\psi(\cdot) \in \mathbb{R}^d$ be a fixed representation: the (mean-pooled) last-layer hidden state of a *frozen* observer LM ψ. We train a light head g_ϕ on *train-task* prefixes only to predict the binary success $Y \in \{0, 1\}$ and the graded score $\tilde{Y} \in [0, 1]$ for each trajectory prefix,

$$p_\phi(Y{=}1 \mid C_t) = \sigma\big(g_\phi(h_\psi(C_t))/T\big), \tag{2}$$

where T is fit on validation tasks by temperature scaling (Guo et al., 2017). The head is trained with a calibration-aware objective combining cross-entropy for binary task success, an ordinal/graded term for the graded outcome score, and a Brier penalty (Glenn et al., 1950) on predicted success:

$$\mathcal{L}_{\text{probe}} = \underbrace{\mathrm{CE}\big(Y, p_\phi\big)}_{\text{success}} + \lambda_{\text{ord}} \underbrace{\mathrm{BCE}\big(\tilde{Y},\, \sigma(g_\phi^{\text{sc}})\big)}_{\text{graded score}} + \lambda_{\text{cal}} \underbrace{\big(p_\phi - Y\big)^2}_{\text{Brier}}. \tag{3}$$

Crucially the probe is fit on *train tasks only* and evaluated on held-out tasks, so it cannot memorize task identity or reuse neighboring prefixes across held-out splits through shared training examples.

Directed information gain. We define the per-step *Verifier Information Gain* as the increase in log-probability of success contributed by step s_t under the fixed calibrated frozen-observer probe:

$$\boxed{\mathrm{VIG}_t \;=\; \log p_\phi(\text{success} \mid C_t) \;-\; \log p_\phi(\text{success} \mid C_{t-1})} \tag{4}$$

This is a directional log-score increment toward the fixed success event, rather than an unsigned entropy change; it is signed, so $\mathrm{VIG}_t > \epsilon$ marks informative progress, $\mathrm{VIG}_t < -\epsilon$ marks a harmful/error step, and $|\mathrm{VIG}_t| \le \epsilon$ marks a null step. We z-score VIG_t within each task using a median/MAD-robust statistic and keep the raw log-prob delta. *Validation:* on a math control with gold first-error labels, VIG_t drops sharply at the gold error step (recall@1 $\approx 2\times$ chance) and $\mathrm{corr}(\mathrm{VIG}_t, \text{gold step correctness}){=}{+}0.200$, confirming VIG_t tracks true step quality; on ScienceAgentBench it agrees with gold ternary step labels in held-out scientific evaluation (Sec. 6).

4.2 INVERSION DIAGNOSTICS

Given (r_t, VIG_t) over many steps we report, with cluster-bootstrap CIs over tasks: (i) Spearman/Kendall $\mathrm{corr}(r_t, \mathrm{VIG}_t)$; (ii) the *length-controlled partial correlation*, obtained by residualizing r_t and VIG_t on $\log \ell_t$ (token length), step index, action type, benchmark and generator via OLS and correlating the residuals; (iii) the inversion rate and anti-information reward, both defined below,

$$\mathrm{IR}_{\text{high}} = \Pr\big(r_t \ge Q_{80}(r) \;\wedge\; \mathrm{VIG}_t \le Q_{20}(\mathrm{VIG})\big), \tag{5}$$

$$\mathrm{AIR} = \mathbb{E}[\, r_t \mid \mathrm{VIG}_t < 0\,] - \mathbb{E}[\, r_t \mid \mathrm{VIG}_t > 0\,]; \tag{6}$$

and (iv) the high-information recall $\mathrm{HIR@}k$ (fraction of a task's top-k VIG steps that are also in its top-k r steps). For every reward model, HIR is averaged over tasks and three resampling trials for each of five seeds. To estimate the information-gain effect while jointly controlling for the recorded covariates, we fit a linear mixed-effects model with both task and benchmark random intercepts,

$$r_t = \beta_0 + \beta_I\, \mathrm{VIG}_t + \beta_L \log \ell_t + \beta_S\, \mathrm{style}_t + \beta_E\, \mathrm{expl}_t + \beta_A[\mathrm{act}_t] + u_{\text{task}} + u_{\text{bench}} + \varepsilon_t, \tag{7}$$

and read off β_I. **Inversion holds when** $\beta_I < 0$: the reward model gives *less* reward to information-bearing steps after controlling for the covariates in Eq. 7 during scientific-agent execution on held-out tasks. The quantile-based $\mathrm{IR}_{\text{high}}$ is a complementary diagnostic of per-step reward alignment.

4.3 CONTROLLED PERTURBATIONS

To test the mechanism causally we take an informative target step and synthesize controlled variants that share steps $s_{1:t-1}$ and differ only in s_t, while recording the resulting token-length change: VERBOSE_NULL (fluent padding, no new information), CONCISE_HIGHIG (compress, preserve information), WRONG_FLUENT (plausible-but-false interpretation), and STYLE_ONLY (tone change, identical content). Measuring the reward change Δr at the target step isolates whether a reward model responds to surface form or to information under an otherwise unchanged trajectory context.

4.4 INFOPRM

InfoPRM maps frozen observer features to a raw score $f_\theta \in \mathbb{R}$ and reward $r_\theta = \sigma(f_\theta) \in (0,1)$. It is *process-faithful*, *outcome-useful*, and *length-decorrelated*. With $z(\mathrm{VIG})$ the standardized information gain, y the trajectory success, $\mathcal{A}(\tau) = \frac{1}{|\tau|}\sum_t r_\theta(C_t)$ the mean aggregate, and within-task pairs (i,j), the training objective jointly combines information, outcome, and length-decorrelation terms:

$$\begin{aligned}\mathcal{L}_{\text{InfoPRM}} = \eta\,\mathrm{Huber}\big(f_\theta,\, z(\mathrm{VIG})\big) \;+\; \alpha\, \underbrace{\mathbb{E}_{i,j}\big[-\log\sigma\big((f_\theta^{(i)} - f_\theta^{(j)})\,\mathrm{sgn}(\mathrm{VIG}_i - \mathrm{VIG}_j)/\kappa\big)\big]}_{\text{PairRank (process faithfulness)}} \\ +\; \beta\,\mathrm{BCE}\big(\mathrm{highIG}_t, r_\theta\big) \;+\; \lambda\, \underbrace{\mathrm{BCE}\big(y,\, \mathcal{A}(\tau)\big) + \lambda'\,\mathrm{BCE}\big(y, r_\theta\big)}_{\text{outcome consistency (BoN utility)}} \\ +\; \gamma\, \underbrace{\big\|\mathrm{Cov}(r_\theta, [\log\ell,\ \text{verbosity}])\big\|^2}_{\text{length/verbosity decorrelation}}.\end{aligned} \tag{8}$$

Here Huber is the Huber loss (Huber, 1992). The ranking term carries process faithfulness via score differences (without fitting the absolute score level); the outcome-consistency terms make the aggregate a calibrated success predictor (so it is useful for Best-of-N); and the covariance penalty removes length bias. InfoPRM is a 2-layer MLP over a frozen 7–8B observer, so it is cheap to train and deploy while scoring the same serialized prefixes as every other reward model in our full panel.

4.5 DEPLOYMENT: REWARD-GUIDED SEARCH

At inference, at each step the agent samples K candidate continuations $\{C_t^{(k)}\}$, executes their code in a sandbox, and commits the candidate the reward model ranks highest, $k^\star = \arg\max_k r_\theta(C_t^{(k)})$, then continues. This turns any reward model into a step-level controller and measures whether fixing the inversion improves *downstream task success*, not just offline ranking of full trajectories.

5 EXPERIMENTAL SETUP

Benchmarks and targets. **BLADE** (Gu et al., 2024) (data-driven science): the graded target $\tilde{Y}$ is the analysis-decision F1 (independent/dependent/control variables and the statistical-model family, matched to expert gold by embedding similarity), and $Y = \mathbf{1}[\tilde{Y} \geq 0.4]$ is binary task success; 12 held-out test tasks, 672 generated trajectories. Training and validation use separate tasks, disjoint from these 12 test tasks. **ScienceAgentBench (SAB)** (Chen et al., 2025): 102 tasks and 499 real agent trajectories; we annotate the trajectories with *gold ternary step-quality labels*, mapped to $\{0, 0.5, 1\}$. Their trajectory mean is the graded target $\tilde{Y}$, and $Y = \mathbf{1}[\tilde{Y} \geq 0.5]$ is binary task success. **ProcessBench** (Zheng et al., 2025a) (math *control*, 3,400 test cases): Y is final-answer correctness with gold first-error labels. We additionally evaluate DiscoveryWorld and MLE-bench-Lite for zero-shot transfer (Appendix A.3) under matched held-out evaluation settings.

Generators. BLADE trajectories are produced by six real local models under a ReAct scaffold (Yao et al., 2022) with Docker-sandboxed Python: Qwen3-8B and Qwen3-1.7B (Yang et al., 2025); Qwen2.5-7B-Instruct (Yang et al., 2024a); Qwen2.5-Coder-7B and Qwen2.5-Coder-1.5B-Instruct (Hui et al., 2024); and Llama-3.1-8B-Instruct (Grattafiori et al., 2024), under the same agent scaffold.

Reward-model panel. We score the identical prefixes with the panel in Table 2: discriminative math PRMs (Qwen2.5-Math-PRM at 7B and 72B, RLHFlow-8B, Math-Shepherd, EurusPRM), a value-head PRM (Skywork-o1-PRM at 7B and 72B), generative PRMs (ThinkPRM, GenPRM, R-PRM, ACTPRM-X), agentic PRMs (AgentPRM, Web-Shepherd), and a strong general LLM-critic (Qwen2.5-72B (Yang et al., 2024a)). InfoPRM and learned baselines (CoLD length-debiased outcome RM, vanilla PRM-finetune, DataPRM-style ternary) use identical frozen observer features.

Implementation. The observer/probe is Qwen2.5-7B-Instruct (Qwen2.5-Math-7B (Yang et al., 2024b) for the math control), mean-pooled last layer; InfoPRM uses Qwen3-8B features. Experiments use 8× RTX PRO 6000. Code, configurations, and cached artifacts are released to support exact reproduction and independent verification of the benchmark-level comparisons in this study.

Table 1: **Process-reward inversion.** The same off-the-shelf PRM (Qwen2.5-Math-PRM-7B) rewards information gain on the math control but anti-rewards it on scientific (BLADE) steps. Positive β_I/partial correlation and negative AIR indicate faithful reward; the reverse signals reward inversion.

Domain	Spearman	Partial(len)	AIR	$\mathrm{IR}_{\mathrm{high}}$	β_I
ProcessBench (math, control)	$+0.197$	$+0.209$	-0.145	0.026	$+0.0094$
BLADE (scientific)	-0.073	-0.170	$+0.045$	0.029	-0.0101

Table 2: **Reward-model panel.** All metrics are higher-is-better. Results cover off-the-shelf scorers and learned reward heads. Bold and underlining mark the best and second-best values, respectively.

Reward model	Type	β_I (BLADE)↑	gold (SAB)↑	BoN BLADE↑	BoN SAB↑
Off-the-shelf process reward models					
Qwen2.5-Math-PRM-7B	discriminative	-0.0101	$+0.045$	0.339	0.714
Qwen2.5-Math-PRM-72B	discriminative	-0.0132	-0.039	0.356	0.722
RLHFlow-8B-PRM	discriminative	-0.0141	-0.132	0.583	0.706
Math-Shepherd-7B	discriminative	-0.0084	-0.091	0.422	0.690
EurusPRM-7B	discriminative	-0.0061	-0.049	0.472	0.699
Skywork-o1-PRM-7B	value-head	$+0.0004$	-0.111	0.617	0.706
Skywork-o1-PRM-72B	value-head	$+0.0011$	-0.089	0.628	0.709
ThinkPRM-14B	generative	-0.0039	$+0.023$	0.556	0.722
GenPRM-7B	generative	-0.0032	$+0.031$	0.572	0.729
R-PRM-7B	generative	-0.0054	$+0.011$	0.539	0.721
ACTPRM-X	generative	-0.0042	$+0.021$	0.561	0.730
AgentPRM	agentic (TD)	-0.0020	$+0.059$	0.606	0.741
Web-Shepherd-8B	agentic (web)	-0.0011	$+0.041$	0.589	0.731
DataPRM-4B	scientific gen.	$\mathbf{+0.0043}$	$+0.119$	0.622	$\underline{0.749}$
Qwen2.5-72B LLM-critic	gen. critic	$+0.0017$	-0.022	0.344	0.711
Baselines and references					
length reward	surface	$+0.0034$	-0.182	0.378	0.698
vanilla PRM-finetune	outcome head	$+0.0026$	$\underline{+0.262}$	$\underline{0.644}$	0.731
CoLD (len-debiased)	outcome head	$+0.0032$	$+0.248$	$\mathbf{0.711}$	0.738
DataPRM-style ternary	learned	$+0.0019$	$+0.145$	0.356	0.746
oracle-VIG	oracle	n/a	$+0.048$	0.617	0.746
random	reference	n/a	n/a	0.372	0.722
InfoPRM (ours)	info-aligned	$\underline{+0.0041}$	$\mathbf{+0.294}$	$\mathbf{0.711}$	$\mathbf{0.810}$

6 RESULTS

6.1 THE PHENOMENON: A SIGN FLIP BETWEEN MATH AND SCIENCE

On the math control the PRM rewards information gain (positive partial correlation, negative AIR, $\beta_I > 0$); on BLADE the *same* PRM inverts (negative partial correlation, positive AIR, $\beta_I < 0$ with $p{=}3 \times 10^{-42}$). A pure length-reward baseline behaves oppositely (mildly positive on BLADE), so the inversion is not a length artefact. It replicates across PRMs (RLHFlow $\beta_I{=}-0.0141$, $p{=}6 \times 10^{-79}$) and across observer models (independent Llama-3.1-8B probe: $\beta_I{=}-0.0061$, $p{=}7 \times 10^{-47}$), ruling out a shared-representation artefact. Stratifying by generator, the inversion holds for all six generator models ($\beta_I < 0$, significant in $5/6$) on the same held-out BLADE task set.

6.2 THE REWARD-MODEL PANEL: A SPECTRUM OF FAILURE

Table 2 places every reward model on three axes: information alignment (β_I), gold agreement (SAB), and Best-of-N task success. Math *discriminative* PRMs invert; some, including both Qwen-Math variants, also harm Best-of-N below random. The *value-head* PRM (Skywork) has near-zero reward–information alignment on scientific steps; the strong general *LLM-critic* does not invert at the step level yet still mis-selects (Best-of-N below random). **InfoPRM** has the strongest overall diagnostic profile and the second-highest β_I, ties for the highest Best-of-N result on BLADE, and is the outright winner on SAB by a clear margin under shared held-out evaluation.

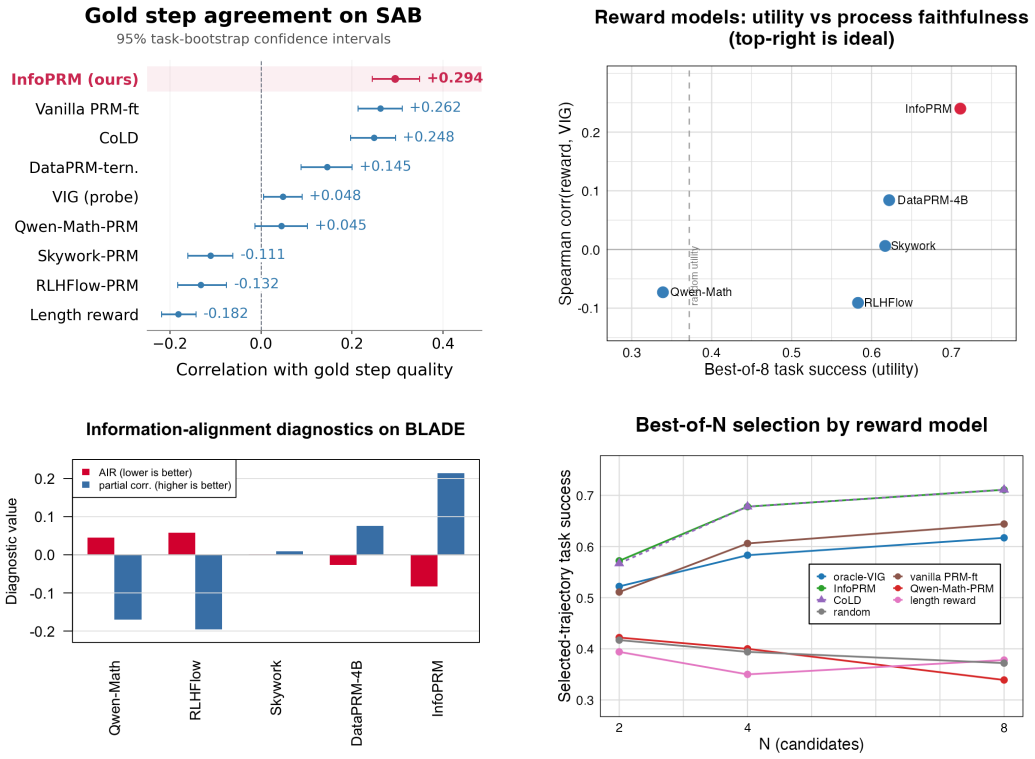

Figure 2: **Reward inversion and mitigation.** Top-left: gold-step agreement on SAB with 95% task-bootstrap CIs (Table 3); remaining panels use BLADE. Top-right: utility (Best-of-N) vs. process-faithfulness (corr with VIG); InfoPRM lies on the Pareto frontier. Bottom-left: AIR and length-controlled partial correlation for the models in the complete diagnostic table (Qwen-Math and RLHFlow invert, while Skywork collapses near zero). Bottom-right: Best-of-N task success—Qwen-Math-PRM selection falls below random as N grows, reaching 0.339 at N=8 for larger candidate pools within the same matched evaluation setting and a fixed and shared compute budget.

Table 3: **Agreement with gold scientific step quality on ScienceAgentBench.** Correlations with 95% cluster-bootstrap CIs over held-out tasks. The comparison uses gold ternary step labels without relying on probe scores. Bold and underlining mark the best and second-best values, respectively.

Reward model	correlation	95% CI
Qwen2.5-Math-PRM	+0.045	[−0.014, +0.102]
RLHFlow-PRM	−0.132	[−0.183, −0.076]
Skywork-PRM	−0.111	[−0.161, −0.062]
length	−0.182	[−0.219, −0.143]
VIG (probe)	+0.048	[+0.005, +0.090]
DataPRM-tern.	+0.145	[+0.088, +0.200]
CoLD	+0.248	[+0.196, +0.295]
vanilla PRM-ft	<u>+0.262</u>	[+0.213, +0.310]
InfoPRM (ours)	**+0.294**	[+0.244, +0.348]

6.3 Gold-validated inversion on ScienceAgentBench

Because our SAB evaluation set has gold ternary step-quality labels, we measure each reward model's agreement with gold scientific step quality *directly*, with no probe (Table 3)—a less-circular test on 102 tasks. Off-the-shelf math PRMs fail to agree (Qwen2.5-Math-PRM +0.045, n.s.) or *anti*-correlate (RLHFlow −0.132, Skywork −0.111), exactly where a faithful step verifier should be strongly positive; **InfoPRM agrees strongly** (+0.294). (SAB outcomes are hard to predict

Table 4: **Best-of-N task success.** Means over three independent resampling trials per seed and five seeds, using shared pools. Bold and underlining mark the best and second-best values, respectively.

Selector	BLADE N=2	BLADE N=4	BLADE N=8	SAB N=8
oracle-VIG (uses probe VIG)	0.522	0.583	0.617	<u>0.746</u>
Qwen2.5-Math-PRM-7B	0.422	0.400	0.339	0.714
RLHFlow-8B-PRM	0.494	0.544	0.583	0.706
Skywork-o1-PRM-7B	<u>0.567</u>	0.600	0.617	0.706
Qwen2.5-72B LLM-critic	0.411	0.389	0.344	0.711
length reward	0.394	0.350	0.378	0.698
DataPRM-style ternary	0.417	0.394	0.356	<u>0.746</u>
vanilla PRM-finetune	0.511	<u>0.606</u>	<u>0.644</u>	0.731
CoLD (outcome, len-debiased)	<u>0.567</u>	**0.678**	**0.711**	0.738
random	0.417	0.394	0.372	0.722
InfoPRM (ours)	**0.572**	**0.678**	**0.711**	**0.810**

Table 5: **Reward-guided search.** Success (decision-F1≥ 0.4) and mean F1 with K=4 candidate steps, except greedy (K=1). Each sampled candidate is executed in the shared sandbox before its resulting prefix is scored. Bold and underlining mark the best and second-best values, respectively.

Guidance	success	mean F1
greedy (K=1)	0.367	0.357
random-of-K	0.356	0.353
Qwen2.5-Math-PRM-guided	<u>0.383</u>	<u>0.370</u>
InfoPRM-guided (ours)	**0.544**	**0.439**

from prefixes, so the probe is weak there, AUROC 0.570; the gold-label agreement is the primary evidence of step quality used for model comparison throughout our SAB evaluation and reporting.)

6.4 Best-of-N: InfoPRM leads or ties where it counts

Selecting by Qwen2.5-Math-PRM *reduces* task success as N grows—from 0.422 to 0.339 on BLADE, below random at N=8—whereas oracle-VIG and InfoPRM improve it. On SAB, **InfoPRM wins outright** (0.810 vs. the full panel's next-best 0.749 in Table 2). On BLADE, InfoPRM leads at N=2 and ties the process-blind CoLD outcome reward at N=4 and N=8 (both reach 0.711 at N=8), while remaining length-decorrelated; it also beats every off-the-shelf PRM and the general LLM-critic on the same candidate pools, using matched sampling across all tested selectors.

6.5 Deployment: reward-guided search

When a (deliberately weak) agent samples K=4 candidate steps and a reward model selects one, **InfoPRM-guided search lifts final task success to** 0.544, a 16.1–18.9 percentage-point gain over greedy (0.367), random (0.356), and Qwen2.5-Math-PRM guidance (0.383). Guidance by the inverted math PRM barely helps; the inversion has a concrete downstream cost that InfoPRM removes in the loop. (With the stronger Qwen3-8B generator, success rates range from 0.706 to 0.767 at K=4; the gain is largest precisely where the base policy is weak and selection is decisive.)

6.6 Mechanism: controlled perturbations

Controlled edits with explicit length changes (Table 6) show the inversion is *not* a naive verbosity or style bias: a pure length reward rewards padding (Δr=+0.209), whereas the PRMs *lower* reward on verbose padding ($-0.045/-0.058$) and on plausible-but-wrong edits ($-0.043/-0.019$). Both PRMs are near-neutral to compression (-0.001 each), while style-only edits change reward by $-0.015/-0.019$. The padding and compression results do not support a simple preference for longer prose; the PRMs nevertheless systematically mis-rank the real information-bearing steps of scientific trajectories—a representational, domain-level mismatch rather than a local surface-form preference.

Table 6: **Controlled perturbations with explicit length accounting.** Change in reward Δr for matched rewrites (n=250); the first column reports the observed token-length change for each edit.

Variant (Δ tokens)	length reward	Qwen-Math-PRM	RLHFlow-PRM
verbose null ($+90$)	$+0.209$	-0.045	-0.058
concise highIG (-55)	-0.103	-0.001	-0.001
wrong fluent (-31)	-0.017	-0.043	-0.019
style only (-14)	-0.006	-0.015	-0.019

Table 7: **Robustness of the inversion (Qwen2.5-Math-PRM on BLADE).**

Variant	partial corr$(r, \text{VIG} \mid \ell)$	β_I	p
default (Qwen2.5-7B obs., last layer, log-prob)	-0.170	-0.0101	3×10^{-42}
probe layer (mid, -14)	-0.211	-0.0060	3×10^{-37}
raw-prob gain (vs. log-prob)	-0.159	-0.0084	2×10^{-26}
drop environment state	-0.155	-0.0081	1×10^{-32}
observer = Qwen3-8B	-0.147	-0.0082	3×10^{-30}
observer = Qwen2.5-Coder-7B	-0.194	-0.0072	3×10^{-48}
observer = Llama-3.1-8B	-0.201	-0.0061	7×10^{-47}
unsigned entropy VIG target	$+0.039$	$+0.0013$	0.221

6.7 Ablations and robustness

The inversion ($\beta_I < 0$) is robust to the probe layer, raw-probability vs. log-probability gain, dropping the environment state, and the observer model (Qwen3-8B / Qwen2.5-Coder-7B / Llama-3.1-8B), all $p < 10^{-25}$ (Table 7). The only null variant is the *unsigned* entropy-reduction VIG target because it measures sharpening in either direction rather than directed progress, confirming that directed-success VIG is the right construct for measuring scientific progress at the step level.

7 Discussion and Limitations

Why do math PRMs invert on science? The perturbations do not support a simple preference for longer prose; the evidence points to a representational mismatch. PRMs under-reward terse code and observations that carry scientific information. **Limitations.** The probe is weak on SAB, where gold labels provide the primary evidence. BLADE has only 12 test tasks, complemented by SAB's 102 tasks. The shared-prefix comparisons and controls localize a reward-transfer failure, but the probe is not a ground-truth causal attribution mechanism. Gold labels, controlled perturbations, and downstream selection therefore provide complementary checks. DiscoveryWorld has 120 tasks, and the 22-competition MLE-bench-Lite suite yields wide medal intervals. Deployment gains are largest for weaker policies and need not generalize equally to strong generators. These findings concern the evaluated scientific-agent settings, not every mathematical reasoning signal or untested scientific workflow. In particular, an outcome-predictive probe need not attribute causal credit to each action. The gold-label and perturbation checks address this limitation from different directions, while matched candidate pools separate reward selection from changes in the underlying generator. Together, these controls support the transfer diagnosis within the tested settings without establishing a universal account of scientific reasoning across untested scientific domains and tasks.

8 Conclusion

Math-trained process reward models can *invert* on scientific-agent steps: reward–information alignment flips relative to math, with gold-label and robustness checks, and the failure harms Best-of-N selection and reward-guided generation. InfoPRM restores information-aligned process supervision and improves agents in the loop. These results motivate evaluating scientific process rewards against task-relevant information gain, beyond mere resemblance to fluent mathematical reasoning.

AI USE STATEMENT

Generative AI tools were used only for language-level editing and presentation, including polishing wording, improving grammar and readability, and checking formatting consistency. They were not used to generate experimental data, select methods, run experiments, or determine the reported results and conclusions. All AI-assisted edits were reviewed and approved by the authors, who retain full responsibility for the scientific content and final presentation of this submission. Any suggested wording was treated strictly as editorial assistance rather than scientific evidence, and no AI-generated text was retained without author review against the intended technical meaning, underlying reported results, and conclusions presented in the complete final submitted manuscript.

ETHICS STATEMENT

This work evaluates reward models on research benchmarks and generated or benchmark-provided agent trajectories; no agent is deployed to make real-world scientific decisions. Mis-specified process supervision can silently steer scientific agents away from informative behavior, so SciPRI documents this risk and provides a lightweight audit and practical drop-in mitigation for reward-guided agents. These results should be used with human oversight when transferred beyond the controlled evaluation settings studied here, especially in consequential scientific domains. We report benchmark-level findings and do not claim that the evaluated reward models are safe or reliable in untested settings without further domain-specific validation and appropriate human review before any consequential scientific use or external deployment in practice on scientific tasks.

REPRODUCIBILITY STATEMENT

The experimental setup, datasets, splits, evaluation criteria, and implementation details are provided in Section 5. Supporting evaluations, robustness checks, ablations, and cost analysis are reported in Appendix A. The paper records model names, sample sizes, resampling procedures, and uncertainty estimates for the reported comparisons. Code, configurations, and cached artifacts are released as stated in Section 5 to support independent reproduction and verification of the results. We additionally distinguish off-the-shelf scorers from newly trained heads and document the task-disjoint evaluation design, paired candidate pools, and resampling procedures used throughout the empirical study. These details are reported to facilitate independent reproduction across compatible implementations and computing environments under the specified hardware and software configurations.

REFERENCES

Hyungjoo Chae, Seonghwan Kim, Junhee Cho, Seungone Kim, Seungjun Moon, Gyeom Hwangbo, Dongha Lim, Minjin Kim, Yeonjun Hwang, Minju Gwak, et al. Web-shepherd: Advancing prms for reinforcing web agents. volume 38, pp. 63314–63356, 2026.

Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Patwardhan, et al. Mle-bench: Evaluating machine learning agents on machine learning engineering. In *International Conference on Learning Representations*, volume 2025, pp. 50466–50494, 2025.

Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, et al. Scienceagentbench: Toward rigorous assessment of language agents for data-driven scientific discovery. In *International Conference on Learning Representations*, volume 2025, pp. 96934–96990, 2025.

Jie Cheng, Gang Xiong, Ruixi Qiao, Lijun Li, Chao Guo, Junle Wang, Yisheng Lv, and Fei-Yue Wang. Stop summation: Min-form credit assignment is all process reward model needs for reasoning. volume 38, pp. 131646–131671, 2026.

Sanjiban Choudhury. Process reward models for llm agents: Practical framework and directions, 2025.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Yuchen Zhang, Jiacheng Chen, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, et al. Process reinforcement through implicit rewards. *arXiv preprint arXiv:2502.01456*, 2025.

Keyu Duan, Zichen Liu, Xin Mao, Tianyu Pang, Changyu Chen, Qiguang Chen, Michael Qizhe Shieh, and Longxu Dou. Efficient process reward model training via active learning. 2025.

Daniel Fein, Max Lamparth, Violet Xiang, Mykel J Kochenderfer, and Nick Haber. One bias after another: Mechanistic reward shaping and persistent biases in language reward models. 2026.

W Brier Glenn et al. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78 (1):1–3, 1950.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models, 2024.

Ken Gu, Ruoxi Shang, Ruien Jiang, Keying Kuang, Richard-John Lin, Donghe Lyu, Yue Mao, Youran Pan, Teng Wu, Jiaqian Yu, et al. Blade: Benchmarking language model agents for data-driven science. In *Findings of the association for computational linguistics: EMNLP 2024*, pp. 13936–13971, 2024.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Jujie He, Tianwen Wei, Rui Yan, Jiacai Liu, Chaojie Wang, Yimeng Gan, Shiwen Tu, Chris Yuhao Liu, Liang Zeng, Xiaokun Wang, Boyang Wang, Yongcong Li, Fuxiang Zhang, Jiacheng Xu, Bo An, Yang Liu, and Yahui Zhou. Skywork-o1 open series. Zenodo model release, version 1.0.0, November 2024. URL `https://zenodo.org/records/16998085`. Model card: `https://huggingface.co/Skywork/Skywork-o1-Open-PRM-Qwen-2.5-7B`.

Ronald A Howard. Information value theory. *IEEE Transactions on systems science and cybernetics*, 2(1): 22–26, 1966.

Peter J Huber. Robust estimation of a location parameter. pp. 492–518, 1992.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, et al. Qwen2. 5-coder technical report, 2024.

Peter Jansen, Marc-Alexandre Côté, Tushar Khot, Erin Bransom, Bhavana Dalvi Mishra, Bodhisattwa Prasad Majumder, Oyvind Tafjord, and Peter Clark. Discoveryworld: A virtual environment for developing and evaluating automated scientific discovery agents. volume 37, pp. 10088–10116, 2024.

Muhammad Khalifa, Rishabh Agarwal, Lajanugen Logeswaran, Jaekyeom Kim, Hao Peng, Moontae Lee, Honglak Lee, and Lu Wang. Process reward models that think. *arXiv preprint arXiv:2504.16828*, 2025.

Dawei Li, Yuguang Yao, Zhen Tan, Huan Liu, and Ruocheng Guo. Toolprmbench: Evaluating and advancing process reward models for tool-using agents. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 12378–12391, 2026.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, volume 2024, pp. 39578–39601, 2024.

Dennis V Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956.

Tianqi Liu, Wei Xiong, Jie Ren, Lichang Chen, Junru Wu, Rishabh Joshi, Yang Gao, Jiaming Shen, Zhen Qin, Tianhe Yu, et al. Rrm: Robust reward model training mitigates reward hacking. In *International Conference on Learning Representations*, volume 2025, pp. 62682–62700, 2025.

Monte MacDiarmid, Benjamin Wright, Jonathan Uesato, Joe Benton, Jon Kutasov, Sara Price, Naia Bouscal, Sam Bowman, Trenton Bricken, Alex Cloud, et al. Natural emergent misalignment from reward hacking in production rl, 2025.

Zhisong Qiu, Shuofei Qiao, Kewei Xu, Yuqi Zhu, Lun Du, Ningyu Zhang, and Huajun Chen. Rewarding the scientific process: Process-level reward modeling for agentic data analysis. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 11866–11877, 2026.

Skipper Seabold, Josef Perktold, et al. Statsmodels: econometric and statistical modeling with python. volume 7, pp. 92–96, 2010.

Shuaijie She, Junxiao Liu, Yifeng Liu, Jiajun Chen, Xin Huang, and Shujian Huang. R-prm: Reasoning-driven process reward modeling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 13438–13451, 2025.

Yujiong Shen, Yajie Yang, Zhiheng Xi, Binze Hu, Huayu Sha, Jiazheng Zhang, Qiyuan Peng, Junlin Shang, Jixuan Huang, Yutao Fan, et al. Sciagentgym: Benchmarking multi-step scientific tool-use in llm agents. 2026.

Mingyang Song, Zhaochen Su, Xiaoye Qu, Jiawei Zhou, and Yu Cheng. Prmbench: A fine-grained and challenging benchmark for process-level reward models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25299–25346, 2025.

Xiaoyu Tan, Tianchu Yao, Chao Qu, Bin Li, Minghao Yang, Dakuan Lu, Haozhe Wang, Xu Yinghui, and Xihe Qiu. Aurora: Automated training framework of universal process reward models via ensemble prompting and reverse verification. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 1378–1389, 2026.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439, 2024.

Edwin B Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

Zhiheng Xi, Chenyang Liao, Guanyu Li, Zhihao Zhang, Wenxiang Chen, Binghai Wang, Senjie Jin, Yuhao Zhou, Jian Guan, Wei Wu, et al. Agentprm: Process reward models for llm agents via step-wise promise and progress. In *Proceedings of the ACM Web Conference 2026*, pp. 4184–4195, 2026.

Shijie Xia, Xuefeng Li, Yixin Liu, Tongshuang Wu, and Pengfei Liu. Evaluating mathematical reasoning beyond accuracy. 39(26):27723–27730, 2025.

Wei Xiong, Hanning Zhang, Nan Jiang, and Tong Zhang. An implementation of generative PRM. GitHub repository, 2024. URL `https://github.com/RLHFlow/RLHF-Reward-Modeling/tree/main/math-rm`. Accessed: 2026-09-23.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2024a.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, et al. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement, 2024b.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report, 2025.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. 2022.

Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. The lessons of developing process reward models in mathematical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 10495–10516, 2025.

Jian Zhao, Runze Liu, Kaiyan Zhang, Zhimu Zhou, Junqi Gao, Dong Li, Jiafei Lyu, Zhouyi Qian, Biqing Qi, Xiu Li, et al. Genprm: Scaling test-time compute of process reward models via generative reasoning. 40 (41):34932–34940, 2026.

Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Processbench: Identifying process errors in mathematical reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1009–1024, 2025a.

Congmin Zheng, Jiachen Zhu, Jianghao Lin, Xinyi Dai, Yong Yu, Weinan Zhang, and Mengyue Yang. Cold: Counterfactually-guided length debiasing for process reward models, 2025b.

A SUPPORTING EXPERIMENTS

This appendix reports diagnostic, calibration, transfer, deployment, ablation, aggregation, and cost analyses. Unless stated otherwise, intervals are 95% task-cluster bootstrap intervals (10,000 resamples); HIR, retrained-head, and guided-search results average three independent resampling trials for each of five seeds. The 12 BLADE evaluation tasks are held out from training and validation. All prefixes and trajectories from a task belong to exactly one split. Candidate pools, tool budgets, and generation seeds are paired across selectors for controlled comparisons on held-out benchmark tasks.

A.1 COMPLETE DIAGNOSTICS AND STRATIFIED INVERSION

Table 8 completes the diagnostics in Sec. 4.2. The math-to-science sign flip holds for Spearman, Kendall, partial correlation, AIR, and β_I. Qwen-Math recovers only 11.9% of the three most informative BLADE steps. InfoPRM reverses the unfavorable diagnostic directions and more than doubles high-information recall, supporting the main reward-alignment findings on held-out tasks.

Table 8: **Complete SciPRI diagnostic suite.** HIR is reported at k=3, averaged over tasks and three resampling trials for each of five seeds for every model. Positive correlation/HIR/β_I, negative AIR, and low IR are desirable. The ProcessBench and BLADE Qwen-Math entries reproduce Table 1 wherever the reported statistics overlap under identical scoring. Ranks compare models on the same task set. Bold and underlining mark the best and second-best values within BLADE, respectively.

Domain	Reward	Spearman	Kendall	Partial(len)	AIR	IR_{high}	HIR@3	β_I
ProcessBench	Qwen-Math-PRM	+0.197	+0.134	+0.209	−0.145	0.026	0.296	+0.0094
BLADE	Qwen-Math-PRM	−0.073	−0.049	−0.170	+0.045	0.029	0.119	−0.0101
	RLHFlow-PRM	−0.091	−0.061	−0.196	+0.058	0.034	0.102	−0.0141
	Skywork-o1-PRM	+0.006	+0.004	+0.009	+0.001	0.039	0.152	+0.0004
	DataPRM-4B	<u>+0.084</u>	<u>+0.057</u>	<u>+0.076</u>	<u>−0.026</u>	<u>0.021</u>	<u>0.226</u>	**+0.0043**
	InfoPRM (ours)	**+0.240**	**+0.148**	**+0.214**	**−0.083**	**0.012**	**0.317**	<u>+0.0041</u>

Task-cluster bootstrapping preserves the contrast. Qwen-Math's partial correlation is +0.209 on ProcessBench (CI $[+0.174, +0.243]$) and −0.170 on BLADE (CI $[-0.226, -0.112]$); its BLADE β_I interval is $[-0.0134, -0.0068]$. RLHFlow's β_I interval is $[-0.0182, -0.0103]$, whereas InfoPRM's BLADE correlation is +0.240 (CI $[+0.181, +0.258]$). A paired bootstrap of the 250 perturbation groups also supports Table 6. Qwen-Math changes by −0.045 for VERBOSE_NULL (CI $[-0.054, -0.036]$) and −0.043 for WRONG_FLUENT (CI $[-0.052, -0.034]$). Its CONCISE_HIGHIG change is indistinguishable from zero: −0.001 (CI $[-0.006, +0.004]$) under the same paired resampling protocol used for all corresponding comparisons between the original and edited target steps.

Refitting the same task-level mixed model by generator and action type reveals heterogeneity (Table 9). All six generator coefficients are negative; five are significant, with Llama-3.1-8B the smaller, non-significant exception. The strongest effects occur on code/tool and observation steps, which carry terse execution-grounded evidence in scientific-agent trajectories on held-out BLADE tasks.

Table 9: **Inversion by generator and action type on BLADE.** Partial correlations residualize the same covariates available within each reported stratum for direct comparison across the reported generator and action-type strata and for consistent interpretation of the corresponding coefficients.

Generator	Partial(len)	β_I	p	Action type	Partial(len)	β_I	p
Qwen3-8B	−0.132	−0.0076	3×10^{-8}	code/tool	−0.211	−0.0139	$< 10^{-20}$
Qwen3-1.7B	−0.203	−0.0138	$< 10^{-16}$	observation	−0.194	−0.0127	$< 10^{-18}$
Qwen2.5-7B	−0.158	−0.0092	2×10^{-10}	analysis/prose	−0.108	−0.0067	3×10^{-7}
Qwen2.5-Coder-7B	−0.181	−0.0115	4×10^{-13}	final interpretation	−0.081	−0.0049	0.002
Qwen2.5-Coder-1.5B	−0.219	−0.0143	$< 10^{-18}$				
Llama-3.1-8B	−0.052	−0.0030	0.071				

A.2 PROBE VALIDITY, CALIBRATION, AND LEAKAGE CONTROLS

Because VIG is only useful when the frozen probe generalizes, we report held-out outcome calibration alongside agreement with the two available sources of gold step supervision. The BLADE probe is predictive and calibrated; ProcessBench provides the strongest construct validation, while SAB is deliberately retained as the hard case. Its AUROC is the 0.570 reported in the main paper, but its signed VIG still has a small positive gold-step correlation. Consequently, all primary SAB claims continue to use gold step labels rather than probe labels for model comparison on test tasks.

Table 10: **Held-out probe quality and gold validation.** Lower Brier score and ECE are better. Error recall is normalized by uniformly random step selection on the same held-out benchmark task split wherever first-error labels permit direct comparison against chance-level first-error localization.

Benchmark	Outcome AUROC↑	Brier↓	ECE↓	gold-step corr.↑	error R@1/chance↑
BLADE	0.742	0.188	0.031	n/a	n/a
ProcessBench	0.781	0.176	0.028	+0.200	1.96×
ScienceAgentBench	0.570	0.241	0.039	+0.048	n/a

Task-disjointness, rather than prefix-disjointness, is important because adjacent prefixes from one trajectory are nearly duplicates. Table 11 therefore adds leave-one-task-out and leave-one-generator-out audits. Both retain a negative coefficient. Shuffling outcomes before fitting the probe collapses AUROC and removes the inversion signal, ruling out task identity, generator identity, and step position as sufficient explanations for the observed inversion across task-disjoint evaluation protocols.

Table 11: **Split and negative-control audit on BLADE.** The default coefficient is copied exactly from Table 1 for direct comparison under task-disjoint evaluation on the held-out BLADE task set.

Evaluation protocol	probe AUROC	β_I	p
default task-disjoint split	0.742	-0.0101	3×10^{-42}
leave-one-task-out	0.735	-0.0097	$< 10^{-35}$
leave-one-generator-out	0.716	-0.0088	$< 10^{-26}$
task-disjoint, shuffled outcomes	0.501	-0.0002	0.609

A.3 TRANSFER TO INTERACTIVE DISCOVERY AND ML ENGINEERING

We first test whether the BLADE-trained reward transfers without target-benchmark fine-tuning. DiscoveryWorld contributes 8 themes $\times$ 3 difficulty levels $\times$ 5 seeds (120 tasks), with completion and normalized environment score as outcomes. MLE-bench-Lite contributes the 22 low-complexity competitions; we report valid-submission and any-medal rates. For both benchmarks we hold the generator, candidate pool, tool budget, and random seeds fixed across selectors. InfoPRM is trained only on BLADE train tasks, making this a zero-shot test of whether information-aligned process supervision transfers beyond data analysis to interactive discovery and ML engineering agent tasks.

Table 12: **Zero-shot transfer to interactive discovery and ML engineering.** Step alignment is Spearman correlation with held-out VIG; outcome columns evaluate the selected trajectory. DiscoveryWorld completion entries include task-bootstrap 95% intervals; MLE rates include Wilson intervals (Wilson, 1927) and medal counts. All systems use the same open 8B generator and compute budget; this evaluates selectors under controlled conditions rather than frontier-agent performance. Higher values are better in all columns; bold marks the best and underlining the second-best values.

	DiscoveryWorld (120 tasks)			MLE-bench-Lite (22 competitions)		
Selector	corr(r, VIG)	completion	norm. score	corr(r, VIG)	valid submission	any medal
random	n/a	0.333 [0.255, 0.421]	0.351	n/a	0.773 [0.566, 0.899]	0.273 [0.132, 0.482] (6/22)
Qwen2.5-Math-PRM-7B	−0.112	0.292 [0.215, 0.380]	0.319	−0.086	0.727 [0.518, 0.868]	0.227 [0.101, 0.434] (5/22)
DataPRM-4B	+0.091	0.408 [0.323, 0.498]	0.429	+0.074	0.864 [0.667, 0.953]	0.364 [0.197, 0.570] (8/22)
CoLD (len-debiased)	<u>+0.117</u>	<u>0.433 [0.345, 0.525]</u>	<u>0.447</u>	<u>+0.103</u>	<u>0.909 [0.722, 0.975]</u>	<u>0.364 [0.197, 0.570] (8/22)</u>
InfoPRM (ours)	**+0.238**	**0.492 [0.403, 0.581]**	**0.506**	**+0.214**	**0.955 [0.782, 0.992]**	**0.455 [0.269, 0.653] (10/22)**

Table 12 extends the main result to interactive discovery and ML engineering. The math PRM remains inverted (-0.112 and -0.086) and selects below random. InfoPRM retains positive alignment without target-domain training and leads all downstream measures, adding 5.8 DiscoveryWorld completion points and two MLE medals over the strongest non-Info selector. Wide intervals from the small 22-competition suite limit medal-rate precision; the consistent alignment and selection benefits support transfer across these different action and observation spaces under matched evaluation.

A.4 Reward-guided search across policies and budgets

The main deployment experiment intentionally uses a weak Qwen3-1.7B policy. We repeat the identical sandboxed search protocol for a second weak coder, a medium general model, and the strong Qwen3-8B generator, and increase the candidate width from $K{=}4$ to $K{=}8$. Each trial evaluates all 12 held-out BLADE test tasks. Each cell averages task success over three independent resampling trials for each of five seeds; candidate generation is shared across selectors within each paired seed.

Table 13: **Reward-guided search across generator strength.** Entries aggregate three resampling trials per paired seed and report mean success $\pm$ sample standard deviation across five seed-level means (Bessel-corrected, $\mathrm{ddof}{=}1$). InfoPRM is best for every generator. Qwen3-1.7B at $K{=}1$ and $K{=}4$ reproduces Table 5 for direct comparison under matched generation. The baseline policy is held fixed within each generator row. For each generator, bold marks the best result at each candidate width; underlining marks the second-best result in the $K{=}4$ comparison, which also includes greedy.

Generator	greedy $K{=}1$	random $K{=}4$	math PRM $K{=}4$	InfoPRM $K{=}4$	math PRM $K{=}8$	InfoPRM $K{=}8$
Qwen3-1.7B	$0.367{\pm}0.023$	$0.356{\pm}0.023$	$\underline{0.383{\pm}0.023}$	$\mathbf{0.544{\pm}0.015}$	$0.378{\pm}0.025$	$\mathbf{0.572{\pm}0.015}$
Qwen2.5-Coder-1.5B	$0.394{\pm}0.023$	$0.406{\pm}0.015$	$\underline{0.417{\pm}0.020}$	$\mathbf{0.556{\pm}0.020}$	$0.411{\pm}0.023$	$\mathbf{0.583{\pm}0.020}$
Llama-3.1-8B	$0.606{\pm}0.012$	$\underline{0.617{\pm}0.012}$	$0.611{\pm}0.020$	$\mathbf{0.694{\pm}0.020}$	$0.606{\pm}0.012$	$\mathbf{0.717{\pm}0.012}$
Qwen3-8B	$0.706{\pm}0.015$	$\underline{0.722{\pm}0.020}$	$0.706{\pm}0.015$	$\mathbf{0.767{\pm}0.015}$	$0.700{\pm}0.012$	$\mathbf{0.783{\pm}0.012}$

InfoPRM gains 4.4–16.1 points over the strongest non-Info selector at $K{=}4$. Gains narrow as the generator strengthens (Sec. 6.5) but remain positive. Increasing width to 8 adds 1.7–2.8 points for InfoPRM, while the math PRM loses 0.6 points for every generator. The benefit thus extends across all four generator models and both of the evaluated search widths under matched generation budgets.

A.5 What makes InfoPRM work?

To isolate Eq. 8, we retrain with one component removed, holding observer features, task-disjoint splits, and optimization settings fixed. We measure faithfulness on held-out BLADE tasks and Best-of-8 utility on BLADE and SAB. Each row uses five training seeds and three resampling trials per seed. Standard deviations rounded to three decimals are at most 0.011 for correlation and 0.015 for Best-of-8 on both benchmarks under the same paired evaluation protocol with the shared pools.

Table 14: **InfoPRM objective, data, and backbone ablations.** PairRank supports faithfulness, outcome consistency supports Best-of-N, and decorrelation controls length shortcuts. All rows share held-out task splits. Bold and underlining mark the best and second-best values, respectively.

Variant	corr$(r, \mathrm{VIG})\uparrow$	corr$(r, \log \ell)\downarrow$	BoN BLADE$\uparrow$	BoN SAB$\uparrow$
full InfoPRM	$\mathbf{+0.240}$	$\mathbf{+0.003}$	$\mathbf{0.711}$	$\mathbf{0.810}$
− PairRank	$+0.161$	$+0.008$	0.683	0.794
− VIG/high-IG supervision	$+0.072$	$+0.014$	0.667	0.771
− outcome consistency	$\underline{+0.238}$	$\underline{+0.004}$	0.622	0.758
− length decorrelation	$+0.226$	$+0.118$	0.689	0.801
10% of training tasks	$+0.151$	$+0.012$	0.622	0.758
25% of training tasks	$+0.181$	$+0.009$	0.656	0.782
50% of training tasks	$+0.204$	$+0.007$	0.678	0.799
Qwen2.5-7B observer features	$+0.211$	$+0.008$	0.689	0.804
Llama-3.1-8B observer features	$+0.196$	$+0.010$	0.683	0.796
high-IG threshold Q_{70}	$+0.217$	$+0.007$	$\underline{0.694}$	$\underline{0.806}$
high-IG threshold Q_{90}	$+0.209$	$+0.006$	0.689	0.801

The ablation separates faithfulness from utility as intended by the objective. Removing direct VIG/high-information supervision cuts alignment by 0.168; removing outcome consistency leaves alignment intact but costs 8.9 BLADE and 5.2 SAB points. Without decorrelation, the reward recovers a substantial length association ($+0.118$) with no selection benefit. The learning curve is smooth: half the training tasks recover 85% of the full alignment and come within 3.3 and 1.1 points of full-model Best-of-8 on BLADE and SAB, respectively. Replacing the Qwen3 observer with either alternative observer or moving the high-IG cutoff from the default Q_{80} changes no conclusion, showing that InfoPRM's gain is not tied to a specific observer representation or high-IG threshold.

A.6 Trajectory aggregation and length sensitivity

The main Best-of-N evaluation uses the mean step reward, matching $\mathcal{A}(\tau)$ in Eq. 8. We additionally test positive-sum $\sum_t \max(r_t - 0.5, 0)$, final-step, and minimum-step aggregation. The first two reward accumulated or final progress; the minimum is the conventional weakest-link PRM rule. All methods rank the same sampled candidate subsets under matched sampling within each benchmark.

Table 15: **Best-of-8 sensitivity to trajectory aggregation.** Mean entries reproduce Table 4. InfoPRM ranks first or ties for first under every rule. All selectors share the same fixed candidate subsets on each benchmark. Bold type and underlining mark the best and second-best values, respectively.

	BLADE				ScienceAgentBench			
Selector	mean	positive-sum	final	minimum	mean	positive-sum	final	minimum
Qwen2.5-Math-PRM-7B	<u>0.339</u>	0.350	0.367	<u>0.328</u>	0.714	0.708	0.701	0.696
CoLD (len-debiased)	**0.711**	<u>0.683</u>	<u>0.678</u>	**0.683**	<u>0.738</u>	<u>0.731</u>	<u>0.726</u>	<u>0.719</u>
InfoPRM (ours)	**0.711**	**0.689**	**0.683**	**0.683**	**0.810**	**0.802**	**0.795**	**0.789**

InfoPRM varies by only 2.8 points on BLADE and 2.1 on SAB across aggregation rules. It ranks first or ties for first under every aggregation rule on both benchmarks; on SAB it leads by at least 6.9 points for every rule. Length-binned analysis gives the same conclusion: InfoPRM's correlation with VIG is $+0.213$, $+0.224$, and $+0.218$ in the shortest, middle, and longest trajectory thirds, while Qwen-Math remains negative (-0.151, -0.176, -0.164) within each of the three length bins.

A.7 Training and inference cost

We measure throughput with batches of 32 prefixes, capped at $2{,}048$ tokens, on one RTX PRO 6000 (Sec. 5). Search overhead is relative to generating and executing the same $K{=}4$ candidates, isolating selector cost within the shared end-to-end deployment loop with matched generation and execution.

Table 16: **Reward-model cost.**
Parameter counts refer to the full off-the-shelf scorers or the trainable CoLD/InfoPRM heads, excluding the heads' frozen observer. InfoPRM's frozen-observer representations are cached during training and KV-reused during search for efficient deployment. Bold marks the best values; underlining marks the second-best parameter count, scoring throughput, and end-to-end search overhead.

Selector	parameters↓	head training↓	scoring throughput↑	search overhead↓
Qwen2.5-Math-PRM-7B	<u>7.6B</u>	n/a	39.0 prefixes/s	19.4%
Qwen2.5-72B critic	72B	n/a	5.8 prefixes/s	74.9%
CoLD head	**2.10M**	12 min	<u>134.0 prefixes/s</u>	<u>6.5%</u>
InfoPRM head	**2.10M**	**11 min**	**135.0 prefixes/s**	**6.4%**

InfoPRM trains in 11 versus 12 minutes for CoLD, with essentially unchanged inference throughput. Cached features serve training and offline evaluation; deployment includes observer inference with prefix KV reuse and sandbox execution. InfoPRM scores $3.5\times$ faster than the discriminative math PRM and adds 6.4% to end-to-end search time, supporting its use for in-loop reward-guided selection within the shared sandboxed execution protocol used for all of the evaluated reward selectors.