Beyond Pairwise Accuracy: Evaluating Reward Models as Optimization Teachers

Anonymous ACL submission

Abstract

Reward models (RMs) are usually evaluated by pairwise preference accuracy, yet reward-guided optimization uses continuous reward signals rather than only binary preferences. We introduce **Reward-Vector Benchmark (RVB)**, a fixed-candidate diagnostic framework for profiling reward models as optimization teachers. For each prompt, RVB asks an RM to score a shared candidate pool and analyzes the resulting reward vector through pairwise correctness, relative spread, concentration, and top-score margin, with cross-prompt contrast retained as an auxiliary diagnostic. We instantiate RVB on a helpfulness upstream benchmark covering 27 reward models, 354 prompts, 3,184 candidate responses, and 85,968 reward-score rows, and relate these offline profiles to reward-side downstream behavior under a controlled RLOO training pipeline, with multi-horizon, PPO-overlap, and harmfulness offline checks for scope calibration. Across the 27RM upstream profile, reward models exhibit heterogeneous teacher-signal regimes that pairwise accuracy alone cannot capture. On the evaluated downstream intersections, RVB descriptors are associated with reward-side speed, stability, and final teacher-reward outcomes beyond pairwise correctness under fixed training recipes. External and held-out RM stress tests further probe transferability and calibration-boundary failure modes. Overall, these results support reward-vector structure as a useful diagnostic axis for profile-aware reward-model analysis under controlled reward-side training recipes.

1 Introduction

Reward models (RMs) are commonly trained and evaluated through pairwise preference comparisons (Ouyang et al., 2022; Bai et al., 2022; Lambert et al., 2025). In this setting, pairwise accuracy is an important sanity metric: it measures how often an RM assigns a higher score to the preferred response in a binary comparison. However, downstream reward-guided procedures, including RLHF optimization and Best-of-N selection (Gao et al., 2023; Zhou et al., 2025), consume *continuous reward signals*, not only binary wins and losses. This creates a practical evaluation gap: an RM may be directionally accurate in pairwise comparisons while still providing a weak, overly compressed, or overly concentrated signal when its scores are used to guide optimization, a concern related to recent observations that RM accuracy and downstream optimization behavior can diverge (Chen et al., 2024; Kim et al., 2025).

The issue is therefore not whether pairwise correctness matters, but whether it is sufficient for evaluating a reward model as an *optimization teacher* (Razin et al., 2025). Under the same prompt and a fixed candidate pool, two RMs may make similar pairwise decisions but organize their reward scores in very different ways. One model may assign nearly indistinguishable scores to candidates of different quality, another may provide graded separation across candidates, and a third may place disproportionate reward pressure on a narrow subset of top candidates. These models can appear similar under pairwise accuracy, but they define different teacher signals for downstream training. From an optimization perspective, the structure of the reward vector—including relative spread, concentration, and top-score margin—determines how much graded preference information is available beyond the binary win/loss decision.

Figure 1 illustrates this gap. All three reward models preserve the same candidate order, but they allocate reward mass differently across the shared candidate set. Under pairwise accuracy alone, these models would look equivalent; under a reward-vector view, they provide distinct teacher signals for optimization.

This paper introduces **Reward-Vector Benchmark (RVB)**, a fixed-candidate framework for ana-

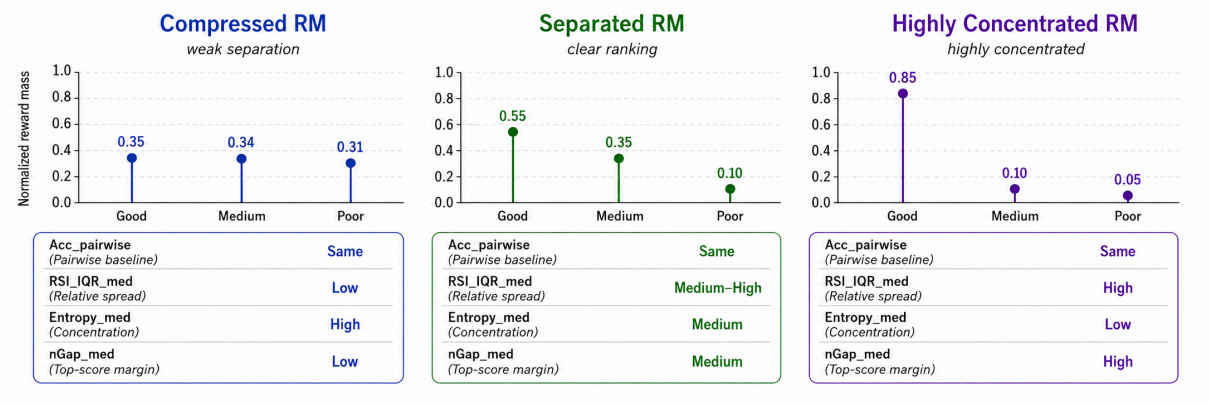

Figure 1: Illustrative synthetic example of reward-vector shapes under a shared candidate set. All three reward models preserve the same candidate order (*Good* > *Medium* > *Poor*), but differ in how they allocate normalized reward mass: compressed, separated, or highly concentrated. This is an illustrative example rather than an empirical RVB result; it shows why pairwise judgment can miss candidate-level teacher-signal structure.

lyzing reward models from this teacher-signal perspective. The name reflects the unit of analysis: for each prompt, RVB studies the reward vector that an RM assigns to a shared candidate pool. This fixed-pool design makes reward-model comparisons more interpretable because score-shape differences are measured over identical prompt-candidate contexts rather than over changing candidate distributions. RVB complements pairwise accuracy with teacher-signal descriptors that capture relative spread, concentration, and top-score margin, while retaining cross-prompt contrast and related quantities as auxiliary diagnostics. Together, these descriptors allow us to ask not only whether an RM selects the preferred response, but also how it shapes the continuous reward landscape consumed by downstream optimization. Unlike calibration-oriented evaluations that ask whether reward scores are probabilistically calibrated, RVB asks how score geometry over a shared candidate pool changes the teacher signal available to reward-guided optimization.

We instantiate RVB on a helpfulness fixed-candidate benchmark derived from RMB (Zhou et al., 2025), covering 27 reward models, 354 prompts, 3,184 candidate responses, and 85,968 reward-score rows. This upstream profile shows substantial heterogeneity in teacher-signal structure across reward models: models with similar pairwise accuracy can occupy different regions of reward-vector geometry. We then examine whether these offline profiles are associated with reward-side downstream behavior under controlled training settings. The main validation centers on a 10-RM RLOO panel, with bounded-horizon, PPO-overlap, and external/held-out RM stress tests used to calibrate scope. In the evaluated downstream intersections, reward-vector descriptors provide diagnostic signal beyond pairwise correctness for reward-side speed, stability, and final teacher-reward outcomes. The supporting analyses further clarify where profile-aware diagnosis transfers and where it remains sensitive to calibration support, reward-model identity, and training recipe.

RVB therefore reframes reward-model evaluation from a binary judging problem to a teacher-signal profiling problem. Instead of asking only whether a reward model assigns the higher score to the preferred response, RVB asks how the model distributes continuous reward information across candidates and whether this structure helps explain reward-side optimization behavior under controlled recipes. The resulting framework identifies a useful diagnostic axis in RM evaluation: reward-vector structure reveals optimization-relevant differences among RMs that pairwise accuracy alone cannot capture. Accordingly, our downstream claims are reward-side and association-level; broader causal, universal-ranking, and human-preference claims are left outside the scope of the current evidence.

Concretely, this paper asks three questions. First, under a shared candidate pool, do reward models with similar pairwise evaluation behavior exhibit distinct reward-vector profiles? Second, do these offline teacher-signal profiles provide downstream-relevant diagnostic signal for reward-side behavior under a fixed RLOO recipe? Third, how do bounded horizon, PPO, and external/held-out RM stress tests refine the transfer potential and calibration limits of profile-based diagnosis?

We summarize our contributions as follows:

- **Teacher-signal formulation**: We formu-

late reward-model evaluation from a teacher-signal perspective, shifting the unit of analysis from binary pairwise judgment alone to prompt-level reward vectors over shared candidate pools.

- **Reward-Vector Benchmark**: We introduce RVB, a fixed-candidate framework for profiling reward-vector structure, and instantiate it on a 27RM helpfulness upstream benchmark showing that pairwise accuracy does not determine reward-vector geometry.
- **Controlled reward-side diagnosis**: We connect RVB profiles to evaluated downstream intersections centered on a 10-RM RLOO panel, with bounded-horizon, PPO-overlap, and stress-test analyses used to probe transfer and calibration limits under controlled reward-side training recipes.

2 Related Work

Reward-model evaluation has traditionally centered on pairwise preference accuracy (Lambert et al., 2025; Zhou et al., 2025). Given a chosen–rejected response pair, a reward model is evaluated by whether it assigns a higher score to the preferred response. This protocol is closely aligned with paired-comparison preference modeling used in RLHF-style alignment pipelines (Ouyang et al., 2022; Bai et al., 2022) and remains a necessary sanity check for detecting clear ordering failures. RewardBench-style evaluations and related robustness benchmarks have therefore been valuable for measuring whether RMs agree with preference labels across diverse settings (Lambert et al., 2025; Wu et al., 2025; Kim et al., 2024; Gureja et al., 2025; Jin et al., 2025; Frick et al., 2024). However, pairwise accuracy mainly evaluates the RM as a binary judge. It does not directly describe how the model distributes continuous reward scores over multiple candidate responses, which is the signal consumed by downstream reward-guided procedures.

Recent RM benchmarks and evaluation protocols have moved toward richer multi-candidate and use-oriented settings, including Best-of-N (BoN) response selection and larger candidate-pool evaluation (Zhou et al., 2025; Malik et al., 2025; Ichihara et al., 2025). These settings better reflect practical uses of reward models in reranking, rejection sampling, preference-data filtering, and post-training (Yuan et al., 2023; He et al., 2024; Yu et al., 2024). Instead of testing only isolated chosen–rejected pairs, they ask whether an RM can select or rank responses from a larger candidate pool. RVB follows this shift from isolated pairs to shared candidate pools, but focuses on a complementary question: under the same prompt and candidate set, how does an RM structure its scores over all candidates? This shifts the unit of analysis from selection correctness alone to prompt-level teacher-signal shape.

A related line of work views reward models not only as preference judges, but also as teachers for optimization (Razin et al., 2025; Chen et al., 2024; Kim et al., 2025). In RLHF-style training, PPO/RLOO-style optimization, Best-of-N selection, and rejection sampling (Ouyang et al., 2022; Gao et al., 2023; Ivison et al., 2024; Zhou et al., 2025), policies consume continuous reward values rather than binary labels alone. From this perspective, score margins, dispersion, calibration, uncertainty, and reward concentration may affect how strongly candidates are distinguished and how stable the reward-guided update becomes (Razin et al., 2025; Leng et al., 2024; Lou et al., 2024; Dorka, 2024). RVB builds on this teacher-oriented view by profiling fixed-pool reward-score structure and examining bounded association-level links with reward-side optimization behavior. Unlike calibration, uncertainty, or leaderboard-oriented evaluations, RVB treats score-shape descriptors as diagnostics tied to the evaluated candidate pool, reward models, and training recipe.

3 RVB Setup and Teacher-Signal Features

RVB evaluates reward models under a fixed candidate-pool protocol. For each prompt x_i, all reward models score the same candidate set

$$\mathcal{Y}_i = \{y_{i1}, \ldots, y_{iK_i}\}.$$

A reward model r therefore produces a prompt-level reward vector

$$\mathbf{s}_{r,i} = \big(s_r(x_i, y_{i1}), \ldots, s_r(x_i, y_{iK_i})\big).$$

This setup makes cross-model comparison more interpretable: differences in teacher-signal structure are measured over the same prompts and candidate responses, rather than over changing prompt or candidate distributions.

The main upstream benchmark is a helpfulness fixed-candidate snapshot with 354 prompts, 3,184 candidate responses, and 85,968 candidate-level

reward scores from 27 reward models. This layer provides the primary substrate for teacher-signal profiling. For each reward model, RVB computes prompt-level descriptors from candidate-level reward vectors and aggregates them into model-level summaries, primarily by taking medians across prompts. The goal is to characterize how different reward models organize continuous reward feedback under the same candidate context.

We separate RVB evidence into three layers: a broad 27RM helpfulness upstream profile, evaluated downstream RLOO/PPO intersections, and an auxiliary harmfulness offline portability check. Upstream claims use the 27RM profile, while downstream claims are restricted to evaluated downstream intersections. A full evidence map is provided in Appendix A.

Teacher-signal descriptors. RVB treats pairwise accuracy as the necessary baseline and then adds reward-vector descriptors that capture how graded reward information is distributed within each fixed candidate pool. The main text focuses on four primary descriptors: `Acc_pairwise`, `RSI_IQR_med`, `Entropy_med`, and `nGap_med`, summarized in Table 1. These descriptors capture binary preference direction, relative reward-vector spread, entropy-based concentration, and the top-score margin between the two highest-scored candidates. Scale-normalized descriptors use an RM-wise robust reference scale, so they compare within-RM reward-vector shape rather than raw reward magnitude. `Entropy_med` is computed from prompt-level reward vectors after a within-prompt softmax transformation and summarizes how reward weight is distributed across candidates. Higher entropy indicates that reward weight is more broadly distributed across the candidate pool, while lower entropy indicates stronger concentration on a smaller subset of candidates. We use this entropy-based quantity as a reward-vector geometry diagnostic, not as a calibrated probability estimate. Auxiliary quantities, including `DCI`, SEI/nGMD variants, lower-level spread primitives, and additional scale-normalized features, are retained as appendix-level diagnostics rather than as the main claim-bearing backbone.

This prompt-centered construction keeps the unit of analysis tied to the shared candidate pool, rather than treating reward scores as independent scalar outputs. It also reduces sensitivity to outlier prompts by aggregating prompt-level descriptors into robust model-level summaries. Raw reward magnitudes are not interpreted as directly comparable absolute utilities across reward models; instead, RVB uses fixed-pool, prompt-level structure to compare how different reward models shape the candidate-level reward landscape. Detailed feature definitions, recomputation details on the fixed data snapshot, auxiliary descriptors, and additional scale-handling discussion are provided in Appendix B.

4 Controlled Downstream Protocol

The upstream RVB profile characterizes how reward models structure fixed-pool reward vectors. The downstream protocol asks whether these offline teacher-signal profiles are associated with reward-side optimization behavior under controlled training settings. We separate the evidential scope of the two layers: the 27RM helpfulness benchmark supports broad offline profiling, while downstream conclusions are restricted to reward models with completed downstream runs.

The main downstream evidence comes from a 10-RM RLOO panel. RLOO enables controlled comparison across reward-model teachers under a shared training recipe, while varying the reward model used as the teacher. For each evaluated RM, the base training setup is kept fixed, so differences in reward-side trajectories can be read as teacher-associated differences under this controlled setup. These trajectories are reward-side optimization signals: they measure how the policy behaves under the reward model used for training, not direct human-preference judgments of final model outputs.

Reward-side targets. We summarize downstream behavior using three reward-side proxies. **Reward-side speed** is measured by normalized reward AUC. **Reward-side stability** is measured by negative tail reward volatility; under this sign convention, larger values indicate lower tail variation and therefore more stable reward-side behavior. **Final teacher-reward outcome** is measured by the final tail reward mean under the selected teacher reward model. Throughout the paper, final teacher-reward outcome refers to reward-side behavior under the training RM, not human-perceived response quality.

Evaluated downstream settings and readouts. The 10-RM panel is the main downstream backbone for association analysis and pseudo-heldout

Role	Feature	What it captures
Pairwise baseline	`Acc_pairwise`	Binary preference direction: whether the RM assigns higher scores to preferred responses in pairwise comparisons.
Relative spread	`RSI_IQR_med`	Scale-normalized reward-vector spread, capturing whether the fixed-pool candidates receive meaningfully separated continuous scores.
Concentration / entropy	`Entropy_med`	Entropy of the within-prompt softmax-transformed reward weights. Higher values indicate more distributed reward information, while lower values indicate stronger concentration on fewer candidates.
Top-score margin	`nGap_med`	Scale-normalized margin between the top-1 and top-2 candidates within a prompt, capturing how sharply the RM isolates the highest-scored response from its nearest competitor.

Table 1: Primary teacher-signal descriptors used in the main RVB analysis. The pairwise baseline measures preference-direction accuracy, while the three continuous descriptors characterize reward-vector structure within a fixed candidate pool.

checking. On this panel, we use two complementary readouts. First, feature–target association analysis relates upstream teacher-signal descriptors to reward-side speed, stability, and final teacher-reward outcome. Second, a pseudo-heldout readout evaluates whether a teacher-signal rule calibrated on the evaluated panel carries information beyond pairwise accuracy alone. We treat this as a bounded diagnostic check rather than as a fully independent held-out generalization test. The Results section reports compact association rows, a pseudo-heldout summary, and SFR/URM stress-test outcomes. Full association counterparts and primitive-feature diagnostics are reported in Appendices C and D.

Supporting checks. We additionally use a six-RM multi-horizon subset at 1,200, 2,400, and 4,800 steps, a three-RM 9,600-step extension, and a three-RM PPO-overlap check. These checks are used to calibrate scope across horizon length and training recipe, rather than to claim full-coverage validation or broad cross-algorithm generalization. Implementation details, representative trajectories, and SFR/URM stress-test identities are provided in Appendix E and the corresponding downstream detail appendices.

5 Main Results

We organize the results as a compact evidence chain. Section 5.1 establishes the upstream premise that pairwise accuracy does not determine reward-vector geometry. Section 5.2 provides a descriptive downstream overlay of profile families. Section 5.3 tests whether teacher-signal features add downstream-relevant signal beyond pairwise accuracy on the main 10-RM RLOO panel. Section 5.4 then uses SFR and URM as RM-level stress tests for transfer potential and calibration boundaries. Supporting horizon, PPO, and harmfulness checks are summarized in Section 5.5. Upstream profiling claims use the helpfulness 27RM fixed-candidate benchmark, while downstream claims remain restricted to evaluated downstream intersections. Throughout this section, final teacher-reward outcome denotes a reward-side proxy under the training RM, not human preference quality.

5.1 Pairwise Accuracy Collapses Distinct Teacher-Signal Regimes

We first test the core upstream premise of RVB: reward models with similar pairwise correctness can still expose different teacher signals once their full reward vectors are examined. Figure 2 maps reward models in a teacher-signal space using `RSI_IQR_med` and `Entropy_med`, with color encoding `Acc_pairwise`. The map shows that pairwise accuracy does not determine full profile geometry: reward models with comparable directional correctness can occupy different spread–entropy regions. This supports the central motivation for RVB: binary preference correctness is necessary, but it collapses reward-vector structure that downstream optimization actually consumes.

We therefore center the main profile interpretation on four primary descriptors: `Acc_pairwise`, `RSI_IQR_med`, `Entropy_med`, and `nGap_med`. These cover pairwise correctness, relative spread, entropy-based concentration, and normalized top-score margin. Auxiliary descriptor families, including `DCI`, SEI/nGMD variants, and additional scale-normalized features, are retained as appendix-level diagnostics rather than as the main claim-bearing

backbone.

Finding. RVB reveals distinct teacher-signal regimes that are invisible to pairwise accuracy alone: under a shared fixed candidate pool, reward models differ not only in whether they prefer the labeled response, but also in how they distribute graded reward information across candidate responses.

5.2 Downstream Overlay of RVB Profiles

We next use the profile-family view as a descriptive downstream overlay, not as a primary validation test. Within the 10-RM downstream panel, the accuracy-oriented / lower-concentration family shows more favorable median reward-side speed and final teacher-reward outcome than the lower-accuracy / compressed-spread family. The lower-accuracy / scale-spread family has no 10-RM downstream coverage and is therefore used only to interpret the upstream profile map. This family-level view provides a coarse bridge to the feature–target analysis in Section 5.3, but it is not treated as standalone evidence of downstream stratification. Representative reward-side trajectories are provided in Appendix F.1.

Finding. Offline RVB profile families provide a descriptive, subset-level overlay that motivates the feature-level beyond-accuracy analysis below.

5.3 Teacher-Signal Features Add Downstream-Relevant Signal Beyond Accuracy

We then test whether RVB features provide downstream-relevant diagnostic signal beyond pairwise accuracy on the main 10-RM RLOO panel. Pairwise accuracy remains a strong and necessary baseline, but it is not sufficient by itself. As shown in Table 2, scale-normalized spread is more strongly associated with reward-side speed and final teacher-reward outcome than accuracy alone when computed on the same fixed data snapshot. Reward-side stability, however, is associated with spread and top-score-margin descriptors in the opposite direction. These patterns indicate that teacher-signal features are target-dependent: features that support faster or higher reward-side improvement need not support lower tail volatility.

To complement the direct association rows, we use a leave-two-out pseudo-heldout diagnostic within the 10-RM panel; protocol details are provided in Appendix C. The calibrated teacher-signal rule reaches an overall tier hit rate of 0.778 and a pairwise directional agreement of 0.778, showing that downstream-relevant reward-side signal cannot be fully reconstructed from `Acc_pairwise` alone within the evaluated RLOO setting.

Appendix C provides the direct appendix counterpart to the selected main-text rows, and Appendix D reports lower-level primitive-feature direct association diagnostics, Pearson counterparts, and leave-one-out ranges.

Finding. RVB profile features provide downstream-relevant diagnostic signal beyond pairwise correctness on the evaluated 10-RM recipe: accuracy remains useful, but spread and top-score-margin-related reward-vector structure add target-dependent reward-side signal under the controlled RLOO setting.

5.4 SFR/URM Stress Tests Expose Transfer Potential and Calibration Boundaries

We further evaluate whether the 10-RM-calibrated teacher-signal rule transfers to stress-test RMs with different identities. SFR is external to the original 27RM upstream panel, with separately computed upstream features. URM is inside the 27RM upstream profile but held out from the 10-RM downstream calibration panel. Thus, SFR probes external transfer, while URM probes a held-out calibration-boundary case.

Under the pre-specified mapping, SFR is a transfer-potential success case with $3/3$ tier hits, while URM is a calibration-boundary failure case with $0/3$ tier hits. A post-hoc recalibrated diagnostic reverses part of this pattern: SFR becomes $0/3$, while URM partially recovers to $2/3$. This contrast shows that teacher-signal profiles can transfer useful downstream information, but that prediction quality is sensitive to calibration support, stress-test RM identity, and rule choice.

Finding. The SFR/URM stress test strengthens the practical interpretation of RVB: profile-based prediction can be useful beyond the calibration panel, but it remains calibration-sensitive rather than universally reliable.

5.5 Supporting Checks Bound the Scope of the Main Claims

Finally, supporting analyses clarify the scope of the main evidence without expanding its claim boundary. The six-RM multi-horizon subset and the three-RM 9,600-step extension reduce concern

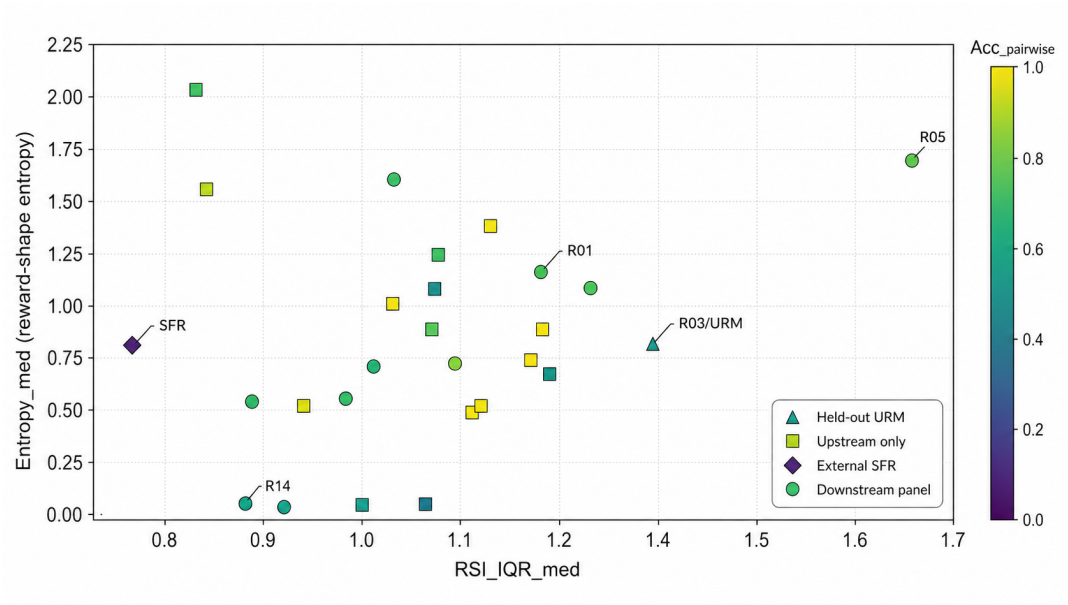

Figure 2: Upstream teacher-signal profile map using `RSI_IQR_med` and reward-shape entropy (`Entropy_med`), with color indicating `Acc_pairwise`. The map shows that pairwise accuracy does not determine reward-vector geometry: RMs in the 10-RM downstream panel, held-out URM, and external SFR occupy distinct regions in the spread–entropy space. This figure is descriptive rather than a leaderboard or causal test. SFR is an external candidate with separately computed upstream features; URM is in the 27RM upstream profile but outside the 10-RM downstream panel.

that the 10-RM patterns are merely artifacts of a single checkpoint. The three-RM PPO overlap check reduces concern that the observed reward-side ordering is a trivial artifact of the RLOO implementation. The harmfulness offline block checks whether the candidate-level scoring and feature-extraction pipeline can run on a safety-oriented split, but it does not support harmfulness downstream optimization claims. Together, these checks provide scope calibration rather than independent full-coverage validation, broad cross-algorithm generalization, or direct human-preference evaluation. Full details are provided in Appendices F.2–G.2.

6 Discussion

From preference judging to teacher-signal profiling. RVB reveals a missing axis in reward-model evaluation: reward-vector teacher-signal structure. The upstream results show that reward models can make similar pairwise decisions while producing different continuous reward landscapes under the same fixed candidate pool. These differences are not reducible to pairwise accuracy alone. Relative spread, entropy-based concentration, and top-score margin expose how much graded information an RM provides and how sharply it isolates top candidates, while auxiliary cross-prompt diagnostics provide additional transparency about consistency across prompts. This is the central reason to evaluate reward models not only as binary preference judges, but also as optimization teachers.

Why score-shape differences matter. The downstream results suggest why these teacher-signal profiles can matter under fixed reward-side training recipes. A compressed reward vector may provide too little graded separation for reward-guided optimization, while an overly sharp or highly concentrated vector may place excessive learning pressure on a narrow subset of candidates. Similarly, entropy helps distinguish whether reward information is distributed across candidates or concentrated on a small subset, which can change the form of reward-side learning pressure. A useful interpretation of the negative associations in Table 2 is therefore not that large top-score margins, high dispersion, or strong top-level separation are inherently harmful. Rather, under a fixed RLOO recipe, these properties may change the effective update scale, alter exploration pressure, or amplify local reward differences in ways that affect reward-side speed, stability, and final teacher-reward outcome.

Profile-aware RM diagnosis and selection. This profile-based view is useful for RM diagnosis, selection, and debugging. If two reward models

Outcome	Feature	Direction	Spearman ρ	LOO stable
Reward-side speed	`Acc_pairwise`	positive	0.732	yes
Reward-side speed	`RSI_IQR_med`	positive	0.891	yes
Final outcome	`Acc_pairwise`	positive	0.750	yes
Final outcome	`RSI_IQR_med`	positive	0.842	yes
Reward-side stability	`Disp_IQR_med`	negative	-0.491	yes
Reward-side stability	`nGap_med`	negative	-0.309	yes

10-RM pseudo-heldout evidence	Speed	Stability	Final outcome	Overall
Tier hit rate	0.844	0.667	0.822	0.778
Pairwise directional agreement	–	–	–	0.778

Table 2: Evidence from the 10-RM downstream panel that teacher-signal features provide downstream-relevant diagnostic signal beyond pairwise accuracy. The upper block reports representative Spearman associations between upstream features and downstream reward-side proxies, including the primary spread descriptor and a primitive spread diagnostic used for the stability readout; the lower block summarizes the pseudo-heldout diagnostic within the same 10-RM panel. Final outcome denotes the final teacher-reward outcome under the training RM, not human preference quality. All results are descriptive, subset-level, and non-causal.

Stress-test RM	Status	Pre-specified	Recalib.	Takeaway
SFR	External stress-test RM outside the original 27-RM panel	$3/3$	$0/3$	Transfer success under the pre-specified rule, but calibration-sensitive after recalibration.
URM	Held-out RM inside the 27-RM profile but outside the 10-RM panel	$0/3$	$2/3$	Calibration-boundary case with partial recovery after recalibration.

Table 3: Compact SFR/URM stress-test summary. The pre-specified and post-hoc recalibrated readouts expose complementary transfer and calibration-boundary behavior. These checks provide stress-test evidence, not universal unseen-RM validation.

have similar pairwise accuracy but occupy different regions of the RVB profile space, downstream differences should not be surprising: the models are supplying different teacher signals to the optimizer. RVB makes these differences measurable before running a full downstream training sweep. The SFR/URM stress tests further show that profile-based reasoning can expose both transfer potential and calibration-boundary cases. In this sense, RVB makes one part of downstream RM selection more explicit: whether a reward model's profile is supported by the calibration regime in which it will be used.

Claim scope. The scope of this interpretation remains controlled. RVB is a diagnostic framework, not a causal explanation of RLHF dynamics or a universal reward-model ranking. Its downstream readouts are reward-side proxies under evaluated training settings, and direct human-preference evaluation remains a complementary requirement for deployment-oriented conclusions. We discuss these boundaries in detail in the Limitations section.

7 Conclusion

This paper introduced Reward-Vector Benchmark (RVB), a fixed-candidate framework for evaluating reward models as optimization teachers. RVB shows that pairwise accuracy can collapse important differences in how reward models distribute continuous reward information across shared candidate pools. Across the 27RM helpfulness upstream profile, reward models exhibit heterogeneous reward-vector structures, and on the evaluated downstream intersections, teacher-signal descriptors provide complementary diagnostic information beyond pairwise correctness. The results support a profile-aware view of RM evaluation: reward models should be compared not only by whether they choose the preferred response, but also by how they shape the reward landscape consumed by optimization. Rather than proposing a universal scalar ranking, RVB offers a controlled framework for diagnosing reward-vector structure and its reward-side downstream implications.

Limitations

RVB is designed as a controlled diagnostic benchmark for studying reward-vector teacher-signal structure, rather than as an exhaustive reward-model leaderboard or a full end-to-end RLHF evaluation suite. The main upstream analysis uses a fixed helpfulness candidate pool covering 27 reward models, 354 prompts, and 3,184 candidate responses. This design improves cross-RM comparability because each reward model scores the same candidate sets, but the resulting profiles may change under different prompt distributions, candidate-generation policies, pool sizes, or application domains.

The downstream evidence is evaluated on controlled intersections with completed training runs. The main validation uses the 10-RM RLOO panel, with additional multi-horizon, long-horizon, and PPO-overlap checks used to calibrate scope. These experiments support association-level claims between upstream teacher-signal descriptors and downstream reward-side behavior under fixed recipes, rather than causal claims about RLHF dynamics in general. Broader coverage across more reward models, seeds, horizons, base policies, and optimization algorithms would further test the generality of these patterns.

The descriptors studied in RVB should be interpreted as diagnostic features of teacher-signal structure. Primary descriptors such as relative spread, concentration, and top-score margin, together with auxiliary diagnostics reported in the appendix, are not intended to provide universal reward-model rankings. Likewise, the downstream targets are reward-side optimization proxies under the selected teacher reward model. Reward-side speed, stability, and final teacher-reward outcome characterize learning behavior under that teacher reward, but they do not replace direct evaluation of final model outputs. The auxiliary harmfulness analysis verifies offline scoring and feature-extraction portability, but does not evaluate harmfulness-oriented downstream optimization. Future work can combine RVB-style diagnostics with human preference labels, high-quality external judges, task-specific utility metrics, and additional safety-oriented downstream settings.

Ethics Statement

This work analyzes existing benchmark artifacts, reward-model scores, and model-generated outputs. We do not deploy systems, collect private user data, or conduct human-subject experiments in this study. The benchmark artifacts and reward models used in our experiments are used for research evaluation and analysis, and we cite their original sources where applicable. We do not redistribute third-party model weights or datasets beyond their permitted access conditions. Released RVB artifacts will be limited to code, derived diagnostic summaries, configuration files, and documentation that can be shared under the corresponding artifact licenses and terms of use. Users of RVB should consult the licenses and intended-use statements of the underlying datasets, reward models, and software packages before reusing them.

RVB is intended to improve transparency in reward-model evaluation by making teacher-signal structure more explicit, not to provide an autonomous certification of model quality or safety. Because reward-model diagnostics can influence model selection and downstream optimization decisions, using RVB-style profiles as standalone rankings may encourage over-reliance on teacher-reward proxies. We therefore recommend pairing reward-side diagnostics with human-centered evaluation, safety-oriented testing, and domain-specific utility checks before deployment.

We used AI assistants to support writing, code/debugging assistance, and organization of experimental artifacts. All scientific claims, experimental results, analysis decisions, and final text were checked and approved by the authors. AI assistants were not used as autonomous evaluators for the reported benchmark scores or downstream results.

References

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, and 12 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *Preprint*, arXiv:2204.05862.

Yanjun Chen, Dawei Zhu, Yirong Sun, Xinghao Chen, Wei Zhang, and Xiaoyu Shen. 2024. The accuracy paradox in RLHF: When better reward models don't yield better language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2980–2989, Miami,

Florida, USA. Association for Computational Linguistics.

Nicolai Dorka. 2024. Quantile regression for distributional reward models in rlhf. *arXiv preprint arXiv:2409.10164*.

Evan Frick, Tianle Li, Connor Chen, Wei-Lin Chiang, Anastasios N Angelopoulos, Jiantao Jiao, Banghua Zhu, Joseph E Gonzalez, and Ion Stoica. 2024. How to evaluate reward models for rlhf. *arXiv preprint arXiv:2410.14872*.

Leo Gao, John Schulman, and Jacob Hilton. 2023. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pages 10835–10866. PMLR.

Srishti Gureja, Lester James Validad Miranda, Shayekh Bin Islam, Rishabh Maheshwary, Drishti Sharma, Gusti Triandi Winata, Nathan Lambert, Sebastian Ruder, Sara Hooker, and Marzieh Fadaee. 2025. M-RewardBench: Evaluating reward models in multilingual settings. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 43–58, Vienna, Austria. Association for Computational Linguistics.

Mingqian He, Yongliang Shen, Wenqi Zhang, Zeqi Tan, and Weiming Lu. 2024. Advancing process verification for large language models via tree-based preference learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2086–2099, Miami, Florida, USA. Association for Computational Linguistics.

Yuki Ichihara, Yuu Jinnai, Tetsuro Morimura, Kenshi Abe, Kaito Ariu, Mitsuki Sakamoto, and Eiji Uchibe. 2025. Evaluation of best-of-n sampling strategies for language model alignment. In *Transactions on Machine Learning Research*.

Hamish Ivison, Yizhong Wang, Jiacheng Liu, Zeqiu Wu, Valentina Pyatkin, Nathan Lambert, Noah A Smith, Yejin Choi, and Hanna Hajishirzi. 2024. Unpacking dpo and ppo: Disentangling best practices for learning from preference feedback. *Advances in neural information processing systems*, 37:36602–36633.

Zhuoran Jin, Hongbang Yuan, Tianyi Men, Pengfei Cao, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. 2025. RAG-RewardBench: Benchmarking reward models in retrieval augmented generation for preference alignment. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17061–17090, Vienna, Austria. Association for Computational Linguistics.

Sunghwan Kim, Dongjin Kang, Taeyoon Kwon, Hyungjoo Chae, Dongha Lee, and Jinyoung Yeo. 2025. Rethinking reward model evaluation through the lens of reward overoptimization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13252–13280, Vienna, Austria. Association for Computational Linguistics.

Sunghwan Kim, Dongjin Kang, Taeyoon Kwon, Hyungjoo Chae, Jungsoo Won, Dongha Lee, and Jinyoung Yeo. 2024. Evaluating robustness of reward models for mathematical reasoning. *arXiv preprint arXiv:2410.01729*.

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. 2025. RewardBench: Evaluating reward models for language modeling. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1755–1797, Albuquerque, New Mexico. Association for Computational Linguistics.

Jixuan Leng, Chengsong Huang, Banghua Zhu, and Jiaxin Huang. 2024. Taming overconfidence in llms: Reward calibration in rlhf. *arXiv preprint arXiv:2410.09724*.

Xingzhou Lou, Dong Yan, Wei Shen, Yuzi Yan, Jian Xie, and Junge Zhang. 2024. Uncertainty-aware reward model: Teaching reward models to know what is unknown. *arXiv preprint arXiv:2410.00847*.

Saumya Malik, Valentina Pyatkin, Sander Land, Jacob Morrison, Noah A Smith, Hannaneh Hajishirzi, and Nathan Lambert. 2025. Rewardbench 2: Advancing reward model evaluation. *arXiv preprint arXiv:2506.01937*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll 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 F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Noam Razin, Zixuan Wang, Hubert Strauss, Stanley Wei, Jason D. Lee, and Sanjeev Arora. 2025. What makes a reward model a good teacher? an optimization perspective. *arXiv preprint arXiv:2503.15477*.

Zhaofeng Wu, Michihiro Yasunaga, Andrew Cohen, Yoon Kim, Asli Celikyilmaz, and Marjan Ghazvininejad. 2025. rewordbench: Benchmarking and improving the robustness of reward models with transformed inputs. *arXiv preprint arXiv:2503.11751*.

Zhuohao Yu, Weizheng Gu, Yidong Wang, Zhengran Zeng, Jindong Wang, Wei Ye, and Shikun Zhang. 2024. Outcome-refining process supervision for code generation. *arXiv e-prints*, pages arXiv–2412.

Hongyi Yuan, Zheng Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. 2023. Rrhf: Rank responses to align language models with human feedback. *Advances in Neural Information Processing Systems*, 36:10935–10950.

Enyu Zhou, Guodong Zheng, Binghai Wang, Zhiheng Xi, Shihan Dou, Rong Bao, Wei Shen, Limao Xiong, Jessica Fan, Yurong Mou, Rui Zheng, Tao Gui, Qi Zhang, and Xuanjing Huang. 2025. Rmb: Comprehensively benchmarking reward models in llm alignment. In *International Conference on Learning Representations*, volume 2025, pages 26543–26589.

A Evidence Map, Data Artifacts, and RM Coverage

This appendix documents the full evidence package for **Reward-Vector Benchmark (RVB)**. To reduce reader burden, Table 4 separates claim-critical evidence, supporting checks, and transparency material. Unless otherwise stated, appendix readouts are descriptive diagnostics under the fixed RVB candidate-pool protocol and evaluated downstream intersections. Supporting checks calibrate scope; they do not expand the main claims beyond the evidence map in Table 5.

Reading priority	Relevant sections	How to use them
Claim-critical evidence	Appendix A Appendix B Appendix C Appendix E Appendix G.1	Evidence scope, RM coverage, teacher-signal definitions, 10-RM association and pseudo-heldout diagnostics, downstream target definitions, and SFR/URM stress-test boundaries.
Supporting checks	Appendix F.1 Appendix F.2 Appendix F.3 Appendix G.2 Appendix G.3	Bounded checks for trajectory illustration, horizon persistence, long-horizon sanity checks, PPO overlap, harmfulness offline portability, and descriptive profile-family organization. These checks contextualize the main evidence but do not enlarge the claim scope.
Transparency only	Appendix A Fig. 3 Appendix D Appendix E	Candidate-pool provenance, lower-level primitive diagnostics, non-redundancy checks, full primitive-feature association tables, artifact paths, and source-level traceability. These sections are not separate leaderboards or independent validation layers.

Table 4: How to read the appendix. The appendix separates claim-critical evidence from supporting checks and transparency material.

RVB uses fixed candidate pools to compare reward models under identical prompt-candidate contexts. For each prompt x_i, all reward models score the same candidate set

$$\mathcal{Y}_i = \{y_{i1}, \ldots, y_{iK_i}\},$$

forming prompt-level reward vectors

$$\mathbf{s}_{r,i} = (s_r(x_i, y_{i1}), \ldots, s_r(x_i, y_{iK_i})).$$

The main upstream helpfulness artifact contains 27 RMs, 354 prompts, 3,184 unique candidates, and 85,968 score rows. The auxiliary harmfulness offline artifact contains an 8RM sample-derived split with 78 prompts and 258 candidate records per RM. Artifact paths and implementation labels are collected in Appendix E.

Fixed-pool candidate construction and analysis scope. RVB separates fixed-pool candidate construction from reward-model scoring and downstream reward-side training. The upstream helpfulness candidate pool was constructed once and then frozen before the analyses reported in this paper. After filtering, this fixed upstream artifact contains 354 prompts, 3,184 retained candidate responses, and 27 reward models, yielding 85,968 score rows in total. The candidate distribution consists of 353 prompts with nine candidates and one prompt with seven valid candidates. Figure 3 documents the candidate-generation template used to produce low-, medium-, and high-quality candidate responses for the fixed pool. All upstream reporting recomputes teacher-signal descriptors on this fixed candidate pool under the current RVB framing—pairwise accuracy, relative spread, entropy-based concentration, and top-score margin—while downstream conclusions remain restricted to the evaluated downstream intersections.

A.1 Reward-model coverage

The full upstream panel defines the broad offline profiling scope, while downstream panels are evaluated intersections with complete downstream runs. Table 6 maps the anonymized RM IDs used throughout the appendix to the underlying reward-model names. These panels are not reward-model leaderboards: inclusion in a downstream panel reflects run availability, compatibility, and intersection constraints rather than an upstream quality judgment.

A.2 Evaluated downstream intersections

Table 7 lists the evaluated downstream intersections used for reward-side analyses. The main 10-RM RLOO panel is the backbone for the association and pseudo-heldout diagnostics, while the smaller multi-horizon, long-horizon, and PPO-overlap subsets provide bounded scope-calibration checks. Because these intersections differ in RM coverage, seed count, horizon, and training recipe, their conclusions are not interchangeable. Downstream claims in the main text are therefore tied to the corresponding evaluated intersection rather

```
You are producing three answers of deliberately different quality tiers (LOW, MEDIUM, HIGH) for the *same* user prompt
to support a Reward Vector benchmark.

Language rule:
- Write in the same language as the user prompt. If unclear, default to English.

Quality definitions (do not disclose to the user):
- HIGH: Fully correct, complete, well-structured. Follows all constraints. Clear reasoning when helpful. No hallucinations.
- MEDIUM: Generally correct but with minor omissions, shallow reasoning, weaker structure, or small style issues.
- LOW: Safe but flawed. Stay on-topic yet intentionally underperform by one or more of: brief/vague, missing key constraints,
shallow or partially incorrect reasoning. Do NOT produce unsafe/offensive content. Do NOT refuse the task.

Stylistic/structural diversity:
- Vary structure, tone, and detail across tiers (list vs. prose; concise vs. detailed).
- Do NOT mention the existence of tiers.
- Rough length guide (flexible for English): LOW 60–120 tokens, MEDIUM 120–200, HIGH 180–300.

Output format:
Return ONLY a JSON object:
{{
 "prompt": "<the prompt text you answered>",
 "answers": [
   {{"quality_tier": "low", "answer": "<text>"}},
   {{"quality_tier": "medium", "answer": "<text>"}},
   {{"quality_tier": "high", "answer": "<text>"}}
 ]
}}

User prompt to answer:
{prompt}
```

Figure 3: Candidate-generation template for the fixed helpfulness candidate pool. The template asks the generator to produce low-, medium-, and high-quality answers to the same source prompt, yielding candidate responses that are later scored by each reward model under the shared fixed-pool protocol. The figure is included only as candidate-pool provenance documentation. The reported RVB analyses use the retained candidate responses and reward-model score vectors, and do not treat this generation template as a separate experimental factor, evaluator, or new contribution.

Evidence layer	Setting	RMs	Data scope	Run scope	Use and boundary
Main upstream profile	Helpfulness fixed candidate pool	27	354 prompts / 3,184 candidates / 85,968 scores	N/A	Primary upstream teacher-signal profiling; no downstream optimization claim from this layer alone
Main downstream association	Main 10-RM RLOO downstream panel	10	Evaluated downstream intersection	5 seeds; standard horizon	Main controlled association and pseudo-heldout readout; not full-27RM downstream validation
Multi-horizon check	Six-RM multi-horizon RLOO subset	6	Evaluated downstream intersection	1,200 / 2,400 / 4,800 steps; 3 seeds each	Horizon-persistence check under the fixed RLOO recipe
Long-horizon extension	Three-RM 9,600-step RLOO extension	3	Evaluated downstream intersection	Up to 9,600 steps; 3 seeds	Long-horizon sanity check; small-subset evidence only
PPO overlap check	Three-RM PPO overlap check	3	Evaluated downstream intersection	1,200 steps; 3 seeds	Bounded check against an obvious RLOO-only artifact
Auxiliary harmfulness sanity	Harmfulness offline sample	8	78 prompts / 258 candidates / 2,064 scores	Offline only	Pipeline-portability sanity under a safety-oriented split; no harmfulness downstream claim

Table 5: Full evidence map and claim scope. RVB combines broad upstream teacher-signal profiling with controlled downstream association evidence and supporting sanity checks.

than to the full 27RM upstream panel.

Evaluated setting	Reward-model IDs
Main 10-RM RLOO panel	R01, R04, R05, R07, R08, R09, R13, R14, R17, R27
Six-RM multi-horizon subset	R01, R04, R05, R07, R08, R14
Three-RM 9,600-step extension	R04, R05, R14
Three-RM PPO overlap check	R04, R05, R14

Table 7: Evaluated downstream intersections used in this paper. Full reward-model names are listed in Table 6.

B Teacher-Signal Feature Definitions

Canonical primitive feature computation is implemented in `stages/utils_features.py` and called by `stages/s2_merge_and_features.py`. The main text uses four reader-facing descriptors: `Acc_pairwise`, `RSI_IQR_med`, `Entropy_med`, and

Idx	Reward model	Idx	Reward model
1	CIR-AMS/BTRM_Qwen2_7b_0613	15	Skywork/Skywork-Reward-V2-Llama-3.2-3B
2	HFXM/RAMO-Llama3.1-8B	16	Skywork/Skywork-Reward-V2-Qwen3-1.7B
3	LxzGordon/URM-LLaMa-3.1-8B	17	Skywork/Skywork-Reward-V2-Qwen3-4B
4	OpenAssistant/reward-model-deberta-v3-large-v2	18	Skywork/Skywork-Reward-V2-Qwen3-8B
5	PKU-Alignment/beaver-7b-v2.0-reward	19	allenai/Llama-3.1-8B-Base-RM-RB2
6	RLHFlow/ArmoRM-Llama3-8B-v0.1	20	allenai/Llama-3.1-8B-Instruct-RM-RB2
7	Ray2333/GRM-Llama3-8B-rewardmodel-ft	21	allenai/Llama-3.1-Tulu-3-8B-DPO-RM-RB2
8	Ray2333/GRM-gemma2-2B-rewardmodel-ft	22	allenai/Llama-3.1-Tulu-3-8B-RL-RM-RB2
9	Ray2333/GRM-llama3.2-3B-rewardmodel-ft	23	allenai/Llama-3.1-Tulu-3-8B-SFT-RM-RB2
10	Ray2333/Gemma-2B-rewardmodel-baseline	24	allenai/tulu-v2.5-13b-uf-rm
11	Skywork/Skywork-Critic-Llama-3.1-8B	25	nicolinho/QRM-Gemma-2-27B
12	Skywork/Skywork-Reward-Gemma-2-27B	26	nicolinho/QRM-Llama3.1-8B-v2
13	Skywork/Skywork-Reward-Llama-3.1-8B	27	weqweasdas/RM-Mistral-7B
14	Skywork/Skywork-Reward-Llama-3.1-8B-v0.2		

Table 6: Full upstream 27RM panel extracted from the canonical helpfulness feature snapshot. Later appendix tables use the reward-model IDs defined here unless otherwise specified.

`nGap_med`. Table 8 summarizes the full descriptor inventory used in the appendix.

For a prompt-level score vector $r = (r_1, \ldots, r_n)$, RVB computes raw spread and concentration primitives:

$$\begin{aligned}
\mathrm{Disp_IQR}(r) &= Q_{75}(r) - Q_{25}(r), \\
\mathrm{Disp_STD}(r) &= \mathrm{STD}_{\mathrm{pop}}(r), \\
\mathrm{Margin_top1_top2}(r) &= r_{(1)} - r_{(2)}, \\
\mathrm{PairSpread_GMD}(r) &= \frac{2}{n(n-1)} \sum_{i<j} |r_i - r_j|.
\end{aligned}$$

RVB also computes the raw entropy descriptor $H(\mathrm{softmax}(r))$. Prompt-level features are aggregated primarily by median, which keeps the analysis prompt-centered and reduces sensitivity to outlier prompts.

B.1 Scale-normalized spread and gap descriptors

For each reward model m, the model-specific reference scale $S_{\mathrm{ref},m}$ is computed from all finite raw scores assigned by that RM in the fixed helpfulness candidate-pool snapshot:

$$S_{\mathrm{ref},m} = 1.4826 \cdot \mathrm{median}(|s - \mathrm{median}(s)|)\,.$$

If this value is non-finite or $\leq 10^{-12}$, the implementation applies fallback branches in order:

$$\mathrm{IQR}(s)/1.349, \quad \mathrm{STD}_{\mathrm{pop}}(s), \quad 1.0.$$

Thus, $S_{\mathrm{ref},m}$ is RM-specific, not prompt-specific and not pooled across RMs.

For prompt-level reward scores r assigned by RM m, RVB computes

$$\mathrm{RSI_IQR}(r, m) = \frac{\mathrm{IQR}(r)}{S_{\mathrm{ref},m}}, \tag{1}$$

$$\mathrm{norm_std}(r, m) = \frac{\mathrm{STD}(r)}{S_{\mathrm{ref},m}}, \tag{2}$$

$$\mathrm{nGap}(r, m) = \frac{r_{(1)} - r_{(2)}}{S_{\mathrm{ref},m}}. \tag{3}$$

The main text uses the median aggregations `RSI_IQR_med` and `nGap_med`; related normalized descriptors are appendix diagnostics.

B.2 Auxiliary descriptors

The scale-normalized global mean pairwise difference is

$$\mathrm{nGMD}(r, m) = \frac{\mathrm{GMD}(r)}{S_{\mathrm{ref},m}}, \tag{4}$$

$$\mathrm{GMD}(r) = \frac{2}{n(n-1)} \sum_{i<j} |r_i - r_j|. \tag{5}$$

The main-text entropy descriptor is the raw softmax entropy aggregated by prompt median:

$$\texttt{Entropy_med} = \mathrm{median}_p \left[H(\mathrm{softmax}(r_p)) \right].$$

For the appendix-level `SEI` diagnostic, RVB instead uses normalized prediction entropy:

$$\mathrm{prediction_entropy} = \frac{H(p)}{\log n}, \tag{6}$$

$$\mathrm{SEI} = 1 - \mathrm{prediction_entropy}, \tag{7}$$

$$p = \mathrm{softmax}(r/\tau), \tag{8}$$

$$\tau = \mathrm{IQR}(r)/1.349. \tag{9}$$

Therefore, `Entropy_med` and `SEI` are related concentration descriptors but not numerically interchangeable.

Descriptor	Role	Aggregation	Interpretation
Acc_pairwise	Pairwise baseline	RM-level agreement	Directional preference correctness in pairwise comparisons.
RSI_IQR_med	Primary descriptor	Median over prompts	Scale-normalized interquartile spread; captures graded separation among fixed-pool candidates.
Entropy_med	Primary descriptor	Median over prompts	Within-prompt softmax entropy; higher values indicate more distributed reward mass, while lower values indicate stronger concentration.
nGap_med	Primary descriptor	Median over prompts	Scale-normalized top-1–top-2 gap; captures how sharply the RM separates the highest-scored candidate from its nearest competitor.
Disp_IQR_med, Disp_STD_med	Primitive diagnostics	Median over prompts	Raw fixed-pool spread primitives.
Margin_top1_top2_med, PairSpread_GMD_med	Primitive diagnostics	Median over prompts	Raw top-1–top-2 margin and pair-spread primitives.
SEI, nGMD, DCI	Auxiliary diagnostics	Prompt or model level	Entropy-derived sharpness, scale-normalized pairwise spread, and cross-prompt consistency diagnostics.

Table 8: Compact descriptor inventory. The first four descriptors form the main reader-facing RVB feature set; primitive and auxiliary quantities are retained for appendix-level diagnostics.

At model level, DCI is computed from robust concentration and spread consistency factors across prompts. Let S_{nGMD} and S_{SEI} denote prompt-level sequences of nGMD and SEI for a given RM:

$$D_{\mathrm{nGMD}} = \frac{\mathrm{median}(S_{\mathrm{nGMD}}) + \epsilon}{\max(\mathrm{IQR}(S_{\mathrm{nGMD}}), \delta)}, \tag{10}$$

$$D_{\mathrm{SEI}} = \frac{\mathrm{median}(S_{\mathrm{SEI}}) + \epsilon}{\max(\mathrm{IQR}(S_{\mathrm{SEI}}), \delta)}, \tag{11}$$

$$\mathrm{DCI} = \exp\left(-\frac{\kappa}{D_{\mathrm{nGMD}} + D_{\mathrm{SEI}}}\right). \tag{12}$$

The implementation constants are $\epsilon = 10^{-6}$, $\delta = 10^{-12}$, and $\kappa = 2.0$. Raw RM score magnitudes are not interpreted as directly comparable absolute utilities across reward models; all descriptors are diagnostic summaries under the fixed candidate-pool and aggregation protocol.

C Core Downstream Diagnostics

This section collects the core 10-RM downstream diagnostics behind the main-text beyond-accuracy claim. It relates the main normalized descriptors to lower-level primitive descriptors, summarizes the selected association rows, and records leave-one-out sensitivity checks.

Table 9 provides a compact view of the descriptor comparison used in the main text. For each downstream target, it contrasts the reported accuracy or primitive reference row with the corresponding core normalized descriptor row. The purpose is to clarify where the main RVB descriptors add signal beyond pairwise accuracy or raw score-structure primitives, rather than to introduce a separate metric leaderboard.

Table 10 expands this compact mapping into the four-block association view used to audit descriptor roles. The table separates accuracy-baseline, selected-reference, core-normalized, and entropy-baseline rows, so that the selected main-text associations can be read in context without treating the table as a metric leaderboard.

The pseudo-heldout readout is computed within the same 10-RM downstream panel using leave-two-out RM pairs. For each held-out pair, the teacher-signal rule is calibrated on the remaining RMs and evaluated on the held-out pair, yielding 45 held-out pair evaluations. For each downstream target, directions are aligned so that larger values indicate more favorable reward-side behavior. The tier-hit rate measures whether the rule assigns the held-out pair to the correct relative downstream tier, while pairwise directional agreement measures whether the predicted pairwise ordering matches the observed reward-side ordering. This is small-panel evidence that RVB features retain directional signal beyond Acc_pairwise, not an independent unseen-RM benchmark.

Table 11 reports compact leave-one-out ranges for representative rows. These ranges are used as a sensitivity check for whether the association direction remains stable when one RM is removed from the 10-RM panel.

Target	Feature	Min	Med.	Max
Reward-side speed	Acc_pairwise	0.653	0.730	0.812
Reward-side speed	RSI_IQR_med	0.850	0.875	0.933
Final teacher-reward outcome	Acc_pairwise	0.655	0.736	0.824
Final teacher-reward outcome	RSI_IQR_med	0.783	0.833	0.900
Reward-side stability	Disp_IQR_med	-0.683	-0.483	-0.300
Reward-side stability	nGap_med	-0.517	-0.317	-0.067

Table 11: Compact leave-one-out ranges for representative rows on the 10-RM downstream panel. Direction stability is assessed under single-RM removal.

D Primitive and Auxiliary Diagnostic Tables

This section groups lower-level primitive and auxiliary diagnostics that support transparency. They

Target	Accuracy / primitive row	Core normalized row	Interpretation
Reward-side speed	Acc_pairwise (0.732)	RSI_IQR_med (0.891)	Normalized spread adds signal.
Reward-side stability	Disp_IQR_med (-0.491)	nGap_med (-0.309)	Stability remains target-dependent.
Final teacher-reward outcome	Acc_pairwise (0.750)	RSI_IQR_med (0.842)	Normalized spread adds signal.

Table 9: Compact mapping between accuracy/primitive descriptors and core normalized descriptors on the 10-RM downstream panel. Values are descriptive Spearman associations under the evaluated downstream intersection.

Target	Feature group	Feature	Spearman ρ	LOO sign stable
Reward-side speed	Accuracy baseline	Acc_pairwise	0.732	yes
Reward-side speed	Selected reference row	Acc_pairwise	0.732	yes
Reward-side speed	Core normalized descriptor	RSI_IQR_med	0.891	yes
Reward-side speed	Entropy baseline	prediction_entropy_med	0.709	yes
Reward-side stability	Accuracy baseline	Acc_pairwise	0.262	yes
Reward-side stability	Selected reference row	Disp_IQR_med	-0.491	yes
Reward-side stability	Core normalized descriptor	nGap_med	-0.309	yes
Reward-side stability	Entropy baseline	prediction_entropy_med	0.176	no
Final teacher-reward outcome	Accuracy baseline	Acc_pairwise	0.750	yes
Final teacher-reward outcome	Selected reference row	Acc_pairwise	0.750	yes
Final teacher-reward outcome	Core normalized descriptor	RSI_IQR_med	0.842	yes
Final teacher-reward outcome	Entropy baseline	prediction_entropy_med	0.661	yes

Table 10: Four-block association summary on the 10-RM downstream panel. The table contextualizes the selected main-text association rows by separating accuracy-baseline, selected-reference, core-normalized, and entropy-baseline diagnostics.

complement the core downstream diagnostics in Appendix C but should not be read as predictor-selection tables.

Table 12 reports the primitive upstream score-structure profile for the full 27RM helpfulness panel. It is included to make the fixed-pool score landscape visible: reward models with similar pairwise behavior can still differ substantially in spread, top-score margin, pair-spread, and entropy.

Table 13 checks whether these primitive score-structure descriptors collapse to pairwise accuracy. The low-to-moderate correlations indicate that pairwise accuracy is informative but does not exhaust the fixed-pool reward-vector structure.

Feature	Spearman ρ with Acc_pairwise
Disp_IQR_med	0.021
Disp_STD_med	-0.085
Margin_top1_top2_med	-0.304
PairSpread_GMD_med	-0.086
Entropy_med	0.264

Table 13: Non-redundancy check between pairwise accuracy and lower-level primitive score-structure descriptors on the 27RM upstream profile.

Tables 14 and 15 provide the full primitive-feature association readouts on the 10-RM downstream panel. These tables are transparency diagnostics: they show the broader association landscape behind the selected rows, but they are not used to select a new primary descriptor set.

Table 16 reports leave-one-out sensitivity for the primitive-feature association rows. It complements the compact core-descriptor leave-one-out summary in Table 11 by showing whether primitive-feature directions are stable under single-RM removal.

Overall, these primitive and auxiliary tables support the main claim that reward-vector structure contains information not captured by pairwise accuracy alone, while remaining descriptive diagnostics under the evaluated 10-RM downstream intersection.

E Downstream Protocol and Reproducibility

This section records downstream target definitions, run labels, and source-level traceability. The controlled RLOO training pipeline is implemented in `clean_rloo_v1/src/run_clean_rloo_v1.py`. The pipeline name and original field names are reported only for reproducibility; the main text uses reader-facing target names.

Table 17 maps the reader-facing downstream targets used in the paper to the implementation fields written by the controlled RLOO pipeline. This mapping clarifies that reward-side speed, stability, and final teacher-reward outcome are fixed-recipe optimization proxies, not direct human-preference metrics.

The tail-window rule used by the implementa-

RM ID	Pairwise acc.	IQR spread	STD spread	Top-1/2 margin	Pair spread	Entropy
R01	0.750	4.846	2.846	0.871	3.539	1.123
R02	0.880	1.996	1.515	0.630	1.872	1.544
R03	0.545	4.466	2.820	2.189	3.533	0.788
R04	0.700	2.585	1.726	0.492	2.145	1.594
R05	0.792	3.806	2.042	0.180	2.480	1.695
R06	0.500	4.305	3.216	1.672	4.057	0.596
R07	0.700	4.938	3.921	1.939	4.745	0.480
R08	0.688	4.996	4.064	2.271	4.957	0.475
R09	0.667	4.086	3.236	1.766	3.862	0.635
R10	1.000	6.488	3.966	1.303	4.951	0.819
R11	0.462	3.623	2.680	1.109	3.307	1.039
R12	0.556	15.207	12.926	5.883	15.802	0.014
R13	0.550	11.907	10.766	6.563	12.681	0.012
R14	0.546	11.430	10.199	5.508	12.208	0.030
R15	1.000	9.524	6.260	2.439	7.812	0.406
R16	1.000	7.051	4.459	1.822	5.496	0.658
R17	0.769	7.845	5.070	1.719	6.212	0.646
R18	1.000	8.333	5.365	2.359	6.677	0.436
R19	1.000	5.706	3.787	1.104	4.683	0.952
R20	1.000	3.620	2.203	0.498	2.733	1.361
R21	0.741	4.162	2.685	0.637	3.317	1.207
R22	0.735	4.086	2.634	0.661	3.277	1.214
R23	0.708	6.890	4.499	1.268	5.625	0.823
R24	1.000	13.695	8.764	2.023	10.794	0.455
R25	0.467	16.570	12.859	5.812	15.880	0.024
R26	0.737	0.737	0.576	0.223	0.717	2.034
R27	0.750	5.378	3.148	1.003	3.917	1.033

Table 12: Primitive upstream score-structure diagnostics for the 27RM helpfulness fixed-candidate profile. Reward-model IDs follow Table 6.

Target	Feature	n	Spearman ρ	Direction
Final teacher-reward outcome	`Acc_pairwise`	10	0.750	positive
Final teacher-reward outcome	`Entropy_med`	10	0.576	positive
Final teacher-reward outcome	`Margin_top1_top2_med`	10	-0.576	negative
Final teacher-reward outcome	`Disp_STD_med`	10	-0.321	negative
Final teacher-reward outcome	`PairSpread_GMD_med`	10	-0.248	negative
Final teacher-reward outcome	`Disp_IQR_med`	10	-0.030	negative
Reward-side speed	`Acc_pairwise`	10	0.732	positive
Reward-side speed	`Margin_top1_top2_med`	10	-0.697	negative
Reward-side speed	`Entropy_med`	10	0.697	positive
Reward-side speed	`Disp_STD_med`	10	-0.455	negative
Reward-side speed	`PairSpread_GMD_med`	10	-0.382	negative
Reward-side speed	`Disp_IQR_med`	10	-0.139	negative
Reward-side stability	`Disp_IQR_med`	10	-0.491	negative
Reward-side stability	`Entropy_med`	10	0.345	positive
Reward-side stability	`Margin_top1_top2_med`	10	-0.345	negative
Reward-side stability	`PairSpread_GMD_med`	10	-0.333	negative
Reward-side stability	`Disp_STD_med`	10	-0.285	negative
Reward-side stability	`Acc_pairwise`	10	0.262	positive

Table 14: Primitive-feature Spearman associations between upstream descriptors and downstream targets on the 10-RM downstream panel.

tion is

$$\texttt{tail_n} = \max(1, \min(100, \lceil 0.2 \times n_points \rceil)).$$

Table 18 lists the run labels corresponding to each evaluated downstream intersection. These labels are intended for source-level traceability and should be read together with the evidence map in Table 5.

Table 19 summarizes the controlled downstream protocol in a reader-facing format. Configuration details such as optimizer settings, generation parameters, and KL regularization are fixed within

Target	Feature	n	Pearson r	Direction
Final teacher-reward outcome	Acc_pairwise	10	0.622	positive
Final teacher-reward outcome	Entropy_med	10	0.573	positive
Final teacher-reward outcome	Margin_top1_top2_med	10	-0.422	negative
Final teacher-reward outcome	Disp_STD_med	10	-0.314	negative
Final teacher-reward outcome	PairSpread_GMD_med	10	-0.312	negative
Final teacher-reward outcome	Disp_IQR_med	10	-0.150	negative
Reward-side speed	Acc_pairwise	10	0.617	positive
Reward-side speed	Entropy_med	10	0.574	positive
Reward-side speed	Margin_top1_top2_med	10	-0.434	negative
Reward-side speed	Disp_STD_med	10	-0.317	negative
Reward-side speed	PairSpread_GMD_med	10	-0.316	negative
Reward-side speed	Disp_IQR_med	10	-0.152	negative
Reward-side stability	Entropy_med	10	0.397	positive
Reward-side stability	Disp_IQR_med	10	-0.347	negative
Reward-side stability	PairSpread_GMD_med	10	-0.307	negative
Reward-side stability	Disp_STD_med	10	-0.299	negative
Reward-side stability	Margin_top1_top2_med	10	-0.280	negative
Reward-side stability	Acc_pairwise	10	0.222	positive

Table 15: Primitive-feature Pearson associations between upstream descriptors and downstream targets on the 10-RM downstream panel.

Target	Feature	LOO min	LOO median	LOO max
Final teacher-reward outcome	Acc_pairwise	0.655	0.736	0.824
Final teacher-reward outcome	Entropy_med	0.500	0.550	0.717
Final teacher-reward outcome	Disp_IQR_med	-0.183	-0.033	0.250
Final teacher-reward outcome	PairSpread_GMD_med	-0.450	-0.217	-0.050
Final teacher-reward outcome	Disp_STD_med	-0.533	-0.283	-0.150
Final teacher-reward outcome	Margin_top1_top2_med	-0.717	-0.550	-0.500
Reward-side speed	Acc_pairwise	0.653	0.730	0.812
Reward-side speed	Entropy_med	0.633	0.692	0.750
Reward-side speed	Disp_IQR_med	-0.333	-0.133	0.033
Reward-side speed	Margin_top1_top2_med	-0.750	-0.692	-0.633
Reward-side speed	PairSpread_GMD_med	-0.550	-0.375	-0.267
Reward-side speed	Disp_STD_med	-0.633	-0.450	-0.350
Reward-side stability	Entropy_med	0.117	0.342	0.567
Reward-side stability	Acc_pairwise	0.059	0.244	0.477
Reward-side stability	Disp_STD_med	-0.483	-0.317	-0.017
Reward-side stability	PairSpread_GMD_med	-0.550	-0.342	-0.083
Reward-side stability	Disp_IQR_med	-0.683	-0.483	-0.300
Reward-side stability	Margin_top1_top2_med	-0.567	-0.342	-0.117

Table 16: Primitive-feature leave-one-out sensitivity ranges on the 10-RM downstream panel, reporting min/median/max correlation after removing one RM at a time. The compact core-descriptor leave-one-out rows are reported in Table 11.

the recorded run packages rather than redefined as separate experimental factors in the paper.

Source artifacts. Table 20 lists the source artifacts used for upstream features, downstream associations, and auxiliary sanity checks. The table is provided for traceability; reuse of third-party datasets, reward models, and software should follow their original licenses, access conditions, and intended-use statements. RVB releases will avoid redistributing third-party artifacts beyond permitted terms and will instead provide code, derived summaries, configuration references, and documentation where appropriate.

Reader-facing target	Implementation field	Definition
Reward-side speed	target_outcome _speed	Normalized reward AUC, implemented as reward_auc_norm.
Reward-side stability	target_outcome _stability	Negative mean tail reward-standard deviation; larger values indicate lower tail volatility.
Final teacher-reward outcome	target_outcome _quality	Final tail reward mean under the selected teacher RM.

Table 17: Mapping between reader-facing downstream targets and implementation fields. These are reward-side optimization proxies under a fixed recipe, not direct human-preference evaluation metrics.

Evaluated setting	Reproducibility label	RMs / seeds
Main RLOO panel	wave2_seed5 _panel10_v1	10 / 5
1,200-step horizon	wave3_long1200 _subset6_v1	6 / 3
2,400-step horizon	wave4_long2400 _subset6_v1	6 / 3
4,800-step horizon	wave5_long4800 _subset6_v1	6 / 3
9,600-step extension	wave6_long9600 _top3_v1	3 / 3
PPO overlap check	wave7_ppo_transfer _subset3_s3 _1200_v1	3 / 3

Table 18: Reproducibility labels for the controlled downstream package. The last column reports the number of evaluated reward models and seeds.

Artifact role	Source path or setting
Main upstream score table	outputs/helpfulness_all_upstream_v1/ scores_merged/helpfulness_all_upstream _scores_merged_candidate_fix_v1.csv
Feature recomputation setting	feature_version= exploratory_v2_evalcore_refscale
Unified upstream feature table	outputs/helpfulness_all_upstream_v1/ analysis/exploratory_v2_evalcore _refscale/model_features_27rm _canonical_plus_exploratory_v1.csv
Downstream target aggregate	clean_rloo_v1/analysis/ panel10_standard_summary_v1.csv
Auxiliary harmfulness artifact	outputs/harmfulness_eval_core/analysis/ harmfulness_sample256 _combined_scoring_check_v1.csv

Table 20: Source artifacts for upstream features, downstream associations, and auxiliary harmfulness sanity checks.

F Supporting Downstream Checks

All checks in this section are scope-calibration diagnostics for the controlled downstream setting. They are read together with the main 10-RM association evidence rather than as independent downstream validation layers.

F.1 Representative reward-side trajectories

The representative trajectory visualization compares R05, R14, and the external stress-test candidate SFR. The curves are centered by each seed's early baseline and truncated to 0–600 training steps, emphasizing within-RM reward-side change rather than raw reward offsets.

F.2 Multi-horizon and long-horizon RLOO checks

Figure 5 visualizes the six-RM multi-horizon subset at 1,200, 2,400, and 4,800 steps. Table 21 provides the per-RM numerical values behind the six-RM horizon plot. Table 22 extends the same readout to the three-RM 9,600-step subset, and Table 23 gives the corresponding mean-over-RMs summary. The long-horizon tables are descriptive sanity checks for horizon persistence, not statistical inference over a broad RM population.

RM	Steps	Seeds	Speed	Stability	Final
R01	1200	3	-2.176	-4.974	-1.771
R01	2400	3	-2.095	-5.130	-4.556
R01	4800	3	-2.101	-4.873	-0.923
R04	1200	3	-3.038	-0.778	-3.060
R04	2400	3	-3.045	-0.770	-2.644
R04	4800	3	-3.039	-0.771	-3.134
R05	1200	3	1.047	-0.951	0.981
R05	2400	3	1.043	-0.950	1.182
R05	4800	3	1.054	-0.909	0.765
R07	1200	3	-5.666	-3.080	-4.453
R07	2400	3	-5.641	-3.051	-5.453
R07	4800	3	-5.585	-3.083	-5.909
R08	1200	3	-6.590	-2.301	-5.171
R08	2400	3	-6.586	-2.223	-6.312
R08	4800	3	-6.545	-2.291	-6.132
R14	1200	3	-5.599	-4.145	-4.030
R14	2400	3	-5.563	-4.064	-5.145
R14	4800	3	-5.469	-4.118	-5.465

Table 21: Per-RM and per-horizon downstream outcomes for the six-RM multi-horizon subset under RLOO. Reward-model IDs follow Table 6.

RM	Steps	Seeds	Speed	Stability	Final
R04	1200	3	-3.038	-0.778	-3.060
R04	2400	3	-3.045	-0.770	-2.644
R04	4800	3	-3.039	-0.771	-3.134
R04	9600	3	-3.022	-0.781	-2.635
R05	1200	3	1.047	-0.951	0.981
R05	2400	3	1.043	-0.950	1.182
R05	4800	3	1.054	-0.909	0.765
R05	9600	3	1.056	-0.960	1.434
R14	1200	3	-5.599	-4.145	-4.030
R14	2400	3	-5.563	-4.064	-5.145
R14	4800	3	-5.469	-4.118	-5.465
R14	9600	3	-5.393	-3.951	-5.583

Table 22: Per-RM and per-horizon outcomes through 9,600 steps for the three-RM long-horizon extension. Reward-model IDs follow Table 6.

Protocol field	Setting in this submission
Base policy / starting model	`meta-llama/Llama-3.2-1B-Instruct`, as recorded in the `clean_rloo_v1` run package.
Training prompt source	Prompt package and run-level manifests recorded in the `clean_rloo_v1` downstream run directories.
Number of RMs (main panel)	10 (R01, R04, R05, R07, R08, R09, R13, R14, R17, R27).
Seeds (main panel)	5 seeds per RM in the main 10-RM panel aggregate.
Standard horizon / steps	1,200 steps (main panel).
Multi-horizon steps	1,200 / 2,400 / 4,800 (six-RM subset) and 9,600 (three-RM extension).
Optimizer / LR / batch-size details	Fixed within each downstream run configuration; detailed values are retained with the run metadata.
Generation settings / KL regularization	Fixed within each downstream run configuration; detailed values are retained with the run metadata.
Teacher RM usage	Each downstream run uses one selected RM as the teacher reward model under a shared training recipe.
Reward aggregation	Per-seed reward-side outcomes are aggregated to RM-level summary statistics in `panel10_standard_summary_v1.csv`.
Reward-side speed formula	`target_outcome_speed = reward_auc_norm`.
Reward-side stability formula	`target_outcome_stability = neg_mean_reward_std_tail`; larger values indicate lower tail volatility.
Final teacher-reward outcome formula	`target_outcome_quality = final_reward_mean`.
Tail-window rule	$\texttt{tail_n} = \max(1, \min(100, \lceil 0.2 \times n_points \rceil))$.
Failed/incomplete-run handling	Completed-run filtering and status accounting are applied before RM-level summarization, as recorded in the downstream run registries.

Table 19: Compact reproducibility summary for the controlled downstream RLOO protocol. Detailed run configurations will be released with the artifact package; fields not expanded in the paper are fixed in the source-locked run metadata rather than varied as experimental factors.

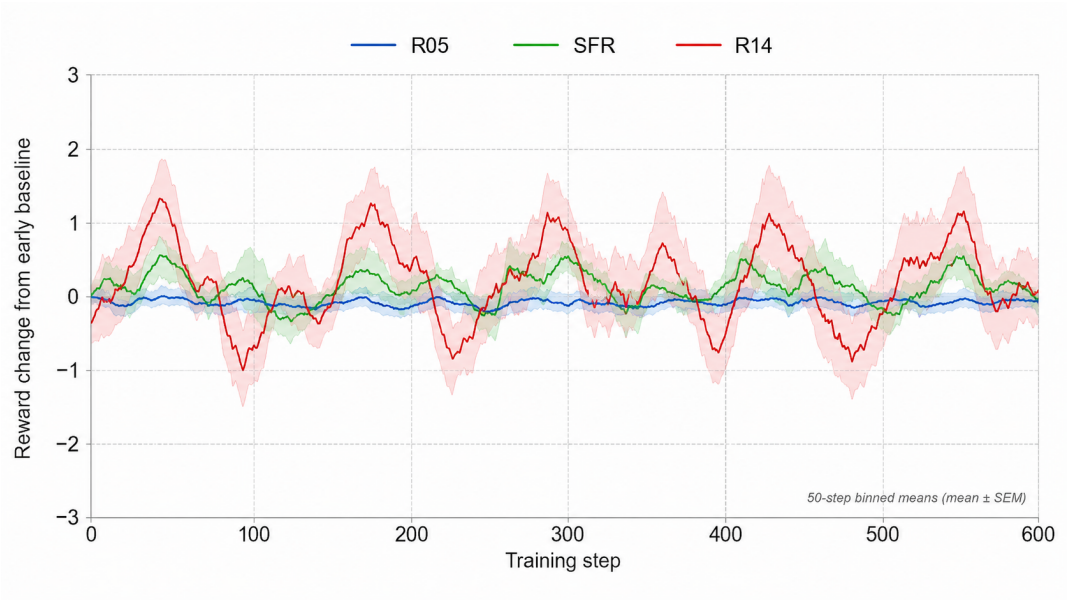

Figure 4: Representative centered RLOO trajectories for R05, R14, and SFR over the first 600 steps. Curves show seed-mean reward-side changes from each seed's early baseline, with SEM bands.

Steps	Mean speed	Mean stability	Mean final
1,200	-2.530	-1.958	-2.036
2,400	-2.521	-1.928	-2.202
4,800	-2.485	-1.933	-2.611
9,600	-2.453	-1.897	-2.261

Table 23: Compact long-horizon summary, reporting means over the three RMs at each horizon. Final denotes final teacher-reward outcome.

F.3 PPO overlap check

The PPO overlap check uses the three reward models that overlap with the RLOO comparison. Its role is to check whether the reward-side ordering is an obvious RLOO-only artifact within this small overlap. Table 24 reports the matched per-RM PPO and RLOO values, while Table 25 summarizes rank-level and linear agreement. Because the overlap contains only three RMs, these tables are scope-calibration diagnostics rather than evidence for broad cross-algorithm generalization.

G Auxiliary Scope and Stress-Test Details

This section collects auxiliary scope checks and stress-test details. SFR/URM directly supports the stress-test interpretation in the main text; harmfulness and profile-family blocks provide offline portability and descriptive organization.

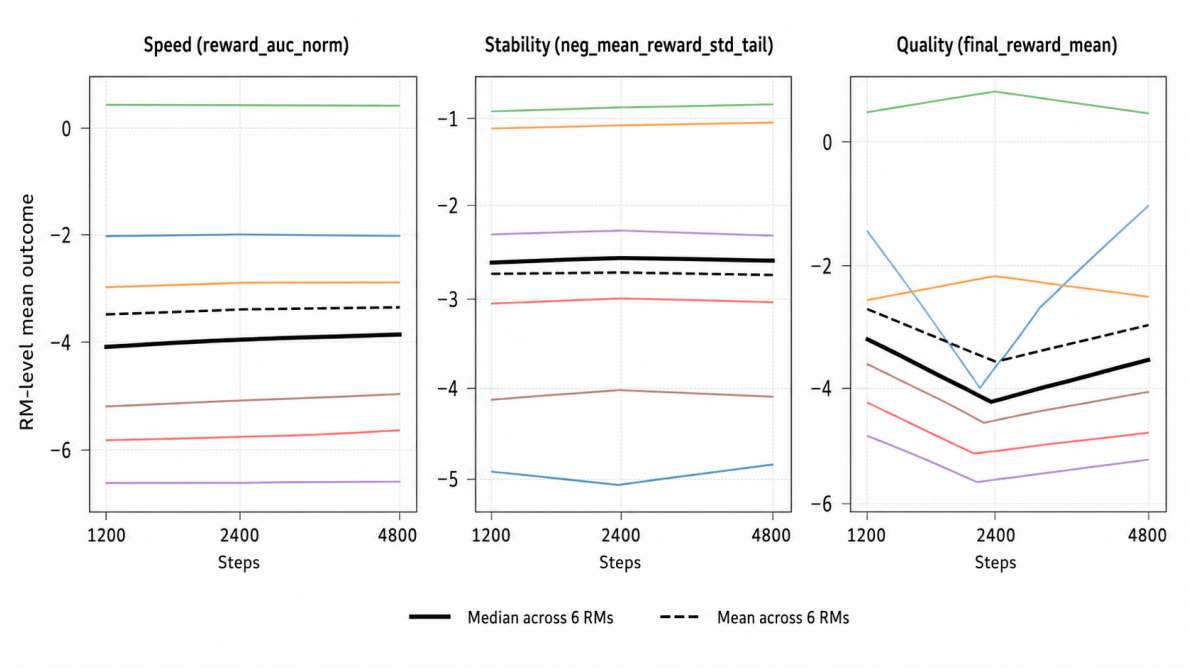

Figure 5: Six-RM multi-horizon reward-side outcomes at 1,200, 2,400, and 4,800 steps. Lines show per-RM outcomes; black summaries show mean and median across the evaluated subset.

RM	Model alias	PPO seeds	RLOO seeds	Speed (PPO)	Speed (RLOO)	Stab. (PPO)	Stab. (RLOO)	Final (PPO)	Final (RLOO)
R04	OpenAssistant DeBERTa-v3	3	3	-3.162	-3.038	-0.355	-0.778	-3.210	-3.060
R05	PKU Beaver-7B	3	3	1.066	1.047	-0.593	-0.951	0.747	0.981
R14	Skywork Llama-3.1-8B v0.2	3	3	-5.676	-5.599	-2.303	-4.145	-6.967	-4.030

Table 24: Per-RM matched PPO and RLOO values on the three-RM overlap. Full reward-model names are defined in Table 6; final denotes final teacher-reward outcome.

Outcome	n	Rank ρ	Pearson r	Top-rank match	Bottom-rank match
Reward-side speed	3	1.000	1.000	1.000	1.000
Reward-side stability	3	1.000	0.998	1.000	1.000
Final teacher-reward outcome	3	1.000	0.948	1.000	1.000

Table 25: Summary metrics for the three-RM PPO overlap check. Top-rank match and bottom-rank match indicate whether PPO and RLOO select the same top and bottom reward model within the overlap.

G.1 SFR/URM Stress-Test Details

SFR is an external stress-test RM outside the original 27RM upstream profile, with separately computed upstream features. URM is included in the 27RM upstream profile but held out from the 10-RM downstream calibration panel. The pre-specified mapping was fixed before inspecting the stress-test outcomes, while the post-hoc recalibrated diagnostic examines calibration sensitivity after observing stress-test behavior. If per-target tier-hit decompositions are included, they should be interpreted only as stress-test diagnostics.

G.2 Harmfulness offline sanity

The harmfulness block verifies that candidate-level scoring and feature extraction run on a safety-oriented split without schema contamination or execution failure. It is offline only and is not paired with harmfulness downstream optimization. Table 26 reports per-RM scoring health, including prompt coverage, nonzero-spread prompts, NaN rows, constant-score checks, and schema validity. Table 27 summarizes the feature-extraction sanity checks. Together, these tables support pipeline portability to the auxiliary harmfulness split, but they do not support a harmfulness downstream optimization claim.

RM ID	Records	Prompts	Multi-cand. prompts	Nonzero-spread prompts	NaN rows	Constant scores	Fields valid
R04	258	78	78	78	0	No	Yes
R05	258	78	78	78	0	No	Yes
R08	258	78	78	78	0	No	Yes
R01	258	78	78	78	0	No	Yes
R07	258	78	78	78	0	No	Yes
R09	258	78	78	78	0	No	Yes
R14	258	78	78	78	0	No	Yes
R17	258	78	78	78	0	No	Yes

Table 26: Harmfulness offline scoring health by RM. Reward-model IDs follow Table 6.

Check	Value
Overall feature-extraction status	Pass
Successful RMs	8
Failed RMs	0
Expected merged score rows	2,064
Feature rows	8
Pairwise-switch format detected	No
Chosen/rejected schema fields valid	Yes
All-NaN feature row present	No

Table 27: Harmfulness feature-extraction sanity summary.

G.3 Teacher-signal profile families

The profile-family block summarizes 27RM offline teacher-signal regimes and their overlap with the evaluated 10-RM downstream panel. Clustering sensitivity compares $k = 3$ and $k = 4$, and the reported configuration uses $k = 3$ based on silhouette and interpretability checks from the same fixed feature table. Family labels are descriptive profile summaries, not discrete quality classes.

Table 28 reports the offline centroid summary for the three teacher-signal profile families. Table 29 overlays those families with the evaluated 10-RM downstream panel. The overlay is included to show coverage and descriptive alignment, not to claim that family assignment determines downstream behavior. Figure 6 visualizes the same primary-descriptor space as a column-standardized descriptive heatmap grouped by these profile families. The heatmap is included to make the offline profile geometry visible; it is not a leaderboard, causal test, or predictor-selection table.

Profile family	RMs	RMs in 10-RM panel	Median pairwise acc.	Median RSI-IQR	Median aux. DCI
F1 (lower-accuracy / compressed-spread)	7	5	0.556	1.038	0.393
F2 (lower-accuracy / scale-spread)	2	0	0.523	1.319	0.863
F3 (accuracy-oriented / lower-concentration)	18	5	0.780	1.156	0.594

Table 28: Compact family centroid summary for teacher-signal profile families. Centroids summarize offline profile regimes across the 27RM upstream panel; DCI is reported as an auxiliary diagnostic rather than a primary descriptor.

Profile family	RMs in 10-RM panel	Median speed	Median stability	Median final outcome
F1 (lower-accuracy / compressed-spread)	5	-5.585	-3.114	-6.792
F2 (lower-accuracy / scale-spread)	0	–	–	–
F3 (accuracy-oriented / lower-concentration)	5	0.048	-1.918	-0.071

Table 29: Downstream overlay by teacher-signal profile family on the 10-RM downstream panel. The overlay is subset-level and descriptive; median final outcome denotes final teacher-reward outcome, not human preference quality.

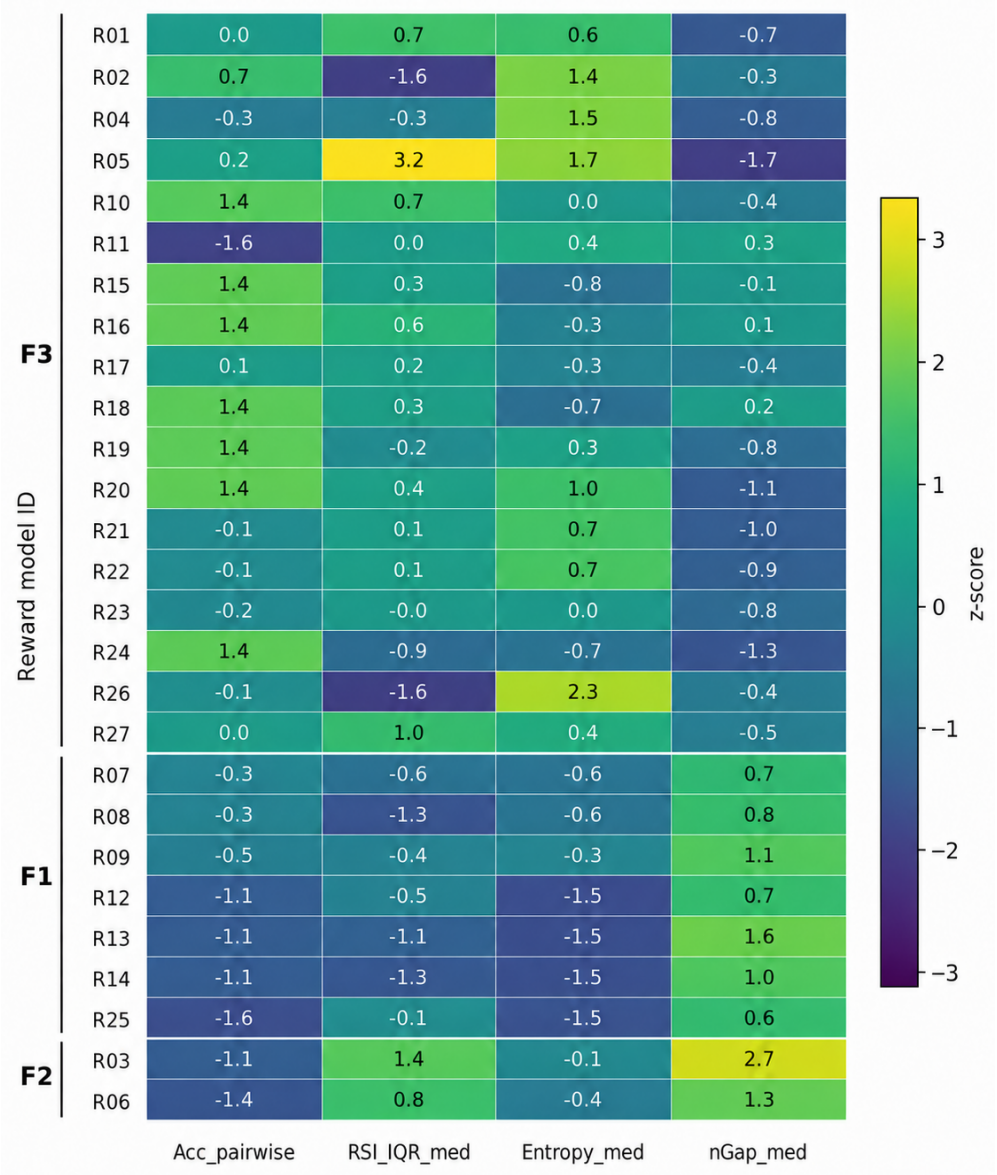

Figure 6: Descriptive heatmap of primary teacher-signal descriptors for the 27RM upstream fixed candidate pool. Rows are reward models, grouped by descriptive profile family, and columns are `Acc_pairwise`, `RSI_IQR_med`, `Entropy_med`, and `nGap_med`. Cell values are column-standardized z-scores across reward models. This panel is descriptive: it is not a leaderboard, not a causal test, and not a predictor-selection table. It supports the main claim that pairwise accuracy does not determine reward-vector profile geometry.