arXiv:2508.03058v1 [cs.LG] 5 Aug 2025

VRPO: Rethinking Value Modeling for Robust RL Training under Noisy Supervision

Dingwei Zhu[1*], Shihan Dou[1*], Zhiheng Xi[1*], Senjie Jin[1], Guoqiang Zhang[1], Jiazheng Zhang[1], Junjie Ye[1], Mingxu Chai[1], Enyu Zhou[1], Ming Zhang[1], Caishuang Huang[1], Yunke Zhang[2], Yuran Wang[2], Tao Gui[1†],

[1]College of Computer Science and Artificial Intelligence, Fudan University
[2]Honor Device Co., Ltd
dwzhu25@m.fudan.edu.cn, tgui@fudan.edu.cn

Abstract

Reinforcement Learning from Human Feedback (RLHF) often suffers from noisy or imperfect reward supervision in real-world settings, which undermines policy stability and generalization. Such noise may cause models to lose attention on key words during advantage estimation. While prior work focuses on reward denoising or filtering poor data, it often overlooks the critical role of the value model in policy optimization. In this work, we show that a strong value model is essential for mitigating noise by absorbing unstable signals and enabling more reliable advantage estimation. We propose VRPO, a value-centric framework for robust PPO training under noisy supervision. VRPO combines two core designs: (1) an auxiliary loss guided by entropy and perplexity from a frozen language model, and (2) a variational information bottleneck. These mechanisms enhance the value model's ability to filter out noise and capture key words from the context during advantage estimation, transforming it from a passive predictor into an active regulator of noise. Experiments on math reasoning, science QA, and multi-turn dialogue, under both rule-based and model-based noisy rewards, show that VRPO consistently outperforms PPO and GRPO baselines. Our findings underscore the often-overlooked importance of the value model in RLHF and offer a principled and practical approach to robust policy optimization in noisy real-world environments.

1 Introduction

Reinforcement Learning (RL) has achieved remarkable success (Perolat et al. 2022; Chen et al. 2022; Bellemare et al. 2020) across a wide range of applications (Zhou et al. 2020; Yue et al. 2024; Caregnato-Neto et al. 2024). However, deploying RL in real-world scenarios often involves noisy or imperfect supervision, particularly when optimization depends on human feedback or learned reward models. This challenge is especially evident in preference-based RL and Reinforcement Learning from Human Feedback (RLHF) (Casper et al. 2023; Zhang et al. 2025), where reward signals are approximate and not directly derived from ground-truth annotations, such as those generated by Generalized Advantage Estimation (GAE) (Schulman et al. 2018).

To address this issue, recent works have proposed robust training techniques by either denoising reward models (Cheng et al. 2025; Miao et al. 2024) or filtering corrupted data before policy updates (Cheng et al. 2024; Wang et al. 2024b). Nevertheless, such methods implicitly assume that reward errors can be corrected during training. In practice, this assumption often fails: reward annotations may be ambiguous, sparse, or intrinsically unreliable (Poiani et al. 2024). Consequently, inaccurate reward signals can propagate through RL training, causing the model to lose attention to key contextual cues during advantage estimation and ultimately harming both performance and convergence.

Surprisingly, while extensive efforts have focused on improving reward modeling, the value model has received little attention as a potential denoising mechanism during training. In this paper, we adopt an information-theoretic perspective (Hung et al. 2023) and propose an alternative path: rather than solely improving reward estimates, we enhance the robustness of PPO by optimizing the value model itself to absorb uncertainty and stabilize training.

We propose VRPO (**Value Model Boosting for Robust Policy Optimization**), a novel framework that incorporates a noise-resilient and semantically aware value model as a core component of robust PPO (Proximal Policy Optimization) training. VRPO introduces two key innovations: (1) auxiliary losses based on entropy and perplexity, and (2) a variational information bottleneck (IB) architecture for robust value modeling. The IB mechanism filters out reward noise by constraining irrelevant information flow, while guidance signals from a frozen language model steer the value predictions to align with meaningful linguistic patterns. These auxiliary objectives further promote partial alignment between the model's internal feature space and the semantic space of language, effectively suppressing spurious noise and preserving essential task-relevant information. In addition, our approach helps alleviate overfitting (Gao, Schulman, and Hilton 2022)and reduces the impact of contextual noise (Wang et al. 2022).

We evaluate VRPO on a wide range of tasks, including mathematical reasoning, scientific question answering, and multi-turn dialogue, under both rule-based and model-based reward settings. Experimental results show that VRPO consistently outperforms both standard PPO and the GRPO

[*] Equal contribution.
[†]Corresponding author.

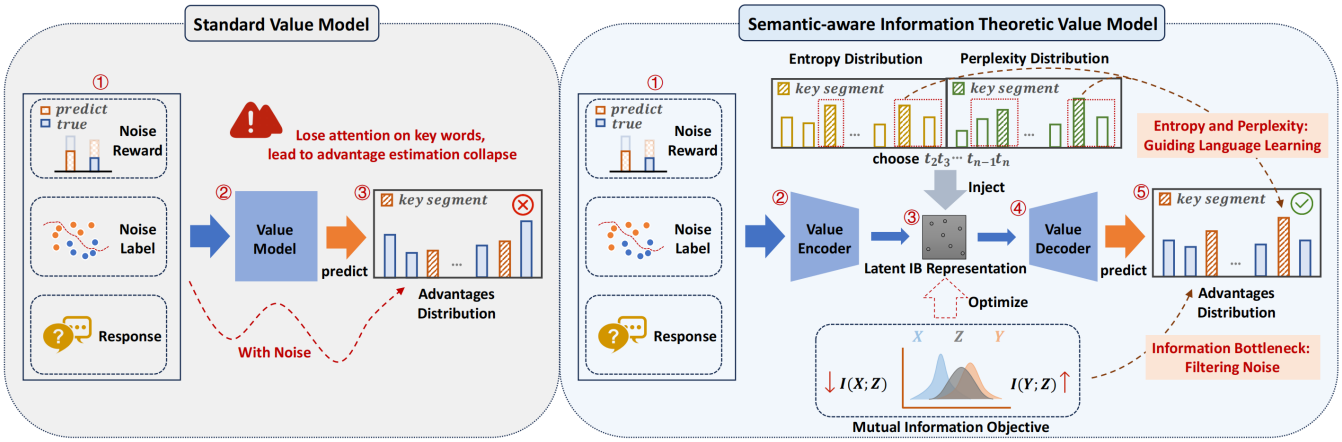

Figure 1: Comparison between the Standard Value Model and our Semantic-aware Information Theoretic Value Model. The standard model fits noisy rewards and labels, leading to unstable advantage estimates. VRPO enhances robustness by guiding learning with entropy and perplexity signals, and filtering noise via a variational information bottleneck, yielding more reliable value predictions and improved PPO stability.

baseline in noisy supervision scenarios. By incorporating information-theoretic regularization and semantic supervision into PPO training, VRPO transforms the value model from a passive estimator into an active, noise-aware component. Through more context-appropriate advantage estimation and key signal extraction, our method enhances learning stability and generalization in challenging environments.

Our contributions are as follows:

- We propose entropy and perplexity-based auxiliary losses to enhance the semantic robustness of the value model under noise.
- We design a variational bottleneck structure to filter irrelevant information during training.
- Empirically, we demonstrate that VRPO improves RL robustness and generalization in noisy RL tasks.

2 Related Work

Robust RL methods primarily focus on improving reward models or filtering noisy data. Reward-oriented approaches build noise-resistant discriminators via conservative gradients (Liang et al. 2024), ensemble models (Wang et al. 2024a), or uncertainty-aware losses (Wu et al. 2024). For example, data-centric methods like RIME (Cheng et al. 2024) use pre-trained denoisers with KL bounds or apply reward consensus to remove inconsistent preferences (Wang et al. 2024a). Our work shifts robustness efforts from reward correction to value model compensation. Through value model optimization, we partially absorb and filter out the noise propagation throughout the training process.

The Information Bottleneck (IB) framework offers a principled way to learn compact (Dai, Zhu, and Wipf 2018; Goyal et al. 2019; Hafner et al. 2019; Ibarz et al. 2018), robust representations by balancing compression and prediction (Wang et al. 2022) through variational approximations (Martini, Abdelaleem, and Nemenman 2024). In RL, IB has enhanced policy or reward robustness under noisy supervision (Miao et al. 2024) but has rarely been applied to value models. We fill this gap by integrating IB regularization into the value function, enabling it to absorb noise during PPO training.

3 Method

3.1 Motivation

In standard Proximal Policy Optimization, policy updates depend on both the reward and value model via advantage estimation. The clipped surrogate objective is:

$$\mathcal{L}_{\text{PPO}} = \mathbf{E}_t \left[\min \left(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t \right) \right] \quad (1)$$

where the advantage $\hat{A}_t$ is typically estimated using Generalized Advantage Estimation (GAE):

$$\hat{A}_t = \delta_t + (\gamma\lambda)\delta_{t+1} + (\gamma\lambda)^2\delta_{t+2} + \cdots, \quad (2)$$

$$\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t) \quad (3)$$

This formulation reveals the tight coupling between the reward r_t and value estimate $V(s)$. When r_t is noisy, its error propagates through $\hat{A}_t$, destabilizing training and harming final policy performance.

We observe that under noisy supervision, the model often **fails to focus on semantically important tokens** in its advantage estimates. As shown in Figure 2 (right), for correct answers, the model tends to allocate higher advantage values toward the end of sequences, resulting in **length hacking** (Chaudhari et al. 2024; Laidlaw, Singhal, and Dragan 2025), where longer responses are rewarded regardless of quality. For incorrect answers, high advantage is mistakenly assigned to irrelevant tokens.

While prior work mostly focuses on denoising rewards or filtering corrupted feedback to address the above issues, we highlight a complementary perspective: enhancing the value model itself to absorb or resist noise, thereby improving training stability.

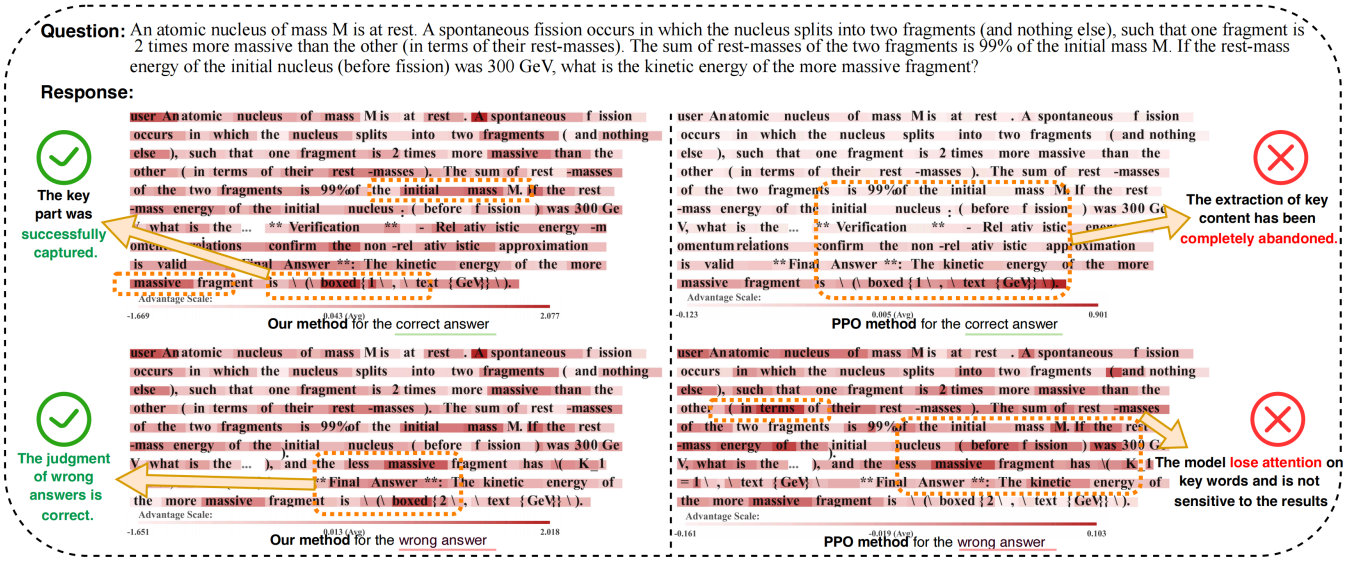

Figure 2: Token-level advantage estimation for the same response across different methods. Our method exhibits sharper **focus on critical reasoning steps**.

3.2 Information-Theoretic Value Modeling

To mitigate reward noise propagation in PPO, the value model must extract return-relevant information while filtering task-irrelevant details. This capability is essential for robust training under ambiguous, sparse, or inconsistent reward signals.

We address this by reformulating value modeling from an information-theoretic perspective. Let the input be a random variable X, the latent representation Z, and the return Y. Assuming Z follows a Gaussian distribution, we define:

$$I_{\text{bottleneck}} = I(X;Z), \quad I_{\text{value}} = I(Z;Y) \tag{4}$$

Here, $I_{\text{bottleneck}}$ measures information retained from the input, while I_{value} quantifies predictive information for returns. The objective is to learn a latent representation that maximizes return relevance while minimizing input redundancy:

$$\max J(\theta) = I(Z;Y) - \beta I(X;Z) \tag{5}$$

where β controls the trade-off between compression and relevance, and θ denotes model parameters.

Since mutual information is intractable in high-dimensional settings, we optimize a variational lower bound using training data $D = \{(x_i, y_i)\}_{i=1}^{N}$:

$$J(\phi,\psi) \geq J_{\text{VLB}}(\phi,\psi) = \mathbf{E}_{(x,y)\sim D}[J_{\text{value}} - \beta J_{\text{bottleneck}}] \tag{6}$$

$$J_{\text{value}} = \mathbf{E}_{z\sim p_\phi(z|x)}[\log q_\psi(y|z)] \tag{7}$$

$$J_{\text{bottleneck}} = \text{KL}[p_\phi(z|x)\|r(z)] \tag{8}$$

where $p_\phi(z|x)$ is the variational encoder, $q_\psi(y|z)$ the return predictor, and $r(z)$ the standard Gaussian prior. Parameters ϕ and ψ correspond to the encoder and decoder.

In practice, we model $p_\phi(z|x)$ as a diagonal Gaussian:

$$z = f_\phi^\mu(x) + f_\phi^\sigma(x)\cdot\epsilon, \quad \epsilon \sim \mathcal{N}(0,I) \tag{9}$$

and use a lightweight MLP q_ψ to predict returns from z. The final training objective becomes:

$$\begin{aligned}\mathcal{L}_{\text{value}} = \mathbf{E}_{(x,y)\sim D}[&-\log q_\psi(z) \\ &+\beta\cdot\text{KL}(p_\phi(z|x)\|r(z))]\end{aligned} \tag{10}$$

This structure encourages the value model to learn compact, reward-relevant latent representations while suppressing irrelevant or noisy input signals.

3.3 Enhancing Semantic Awareness for Noise-Resistant Value Modeling

While the information bottleneck in Section 3.2 helps filter out irrelevant features, it does not directly address semantic misalignment. Under noisy reward supervision, the value model may still lose attention on key words, resulting in prediction errors. To mitigate this, we introduce a semantic-level regularization mechanism that enhances the alignment between the model's feature space and the language space.

Concretely, we incorporate a frozen LM head into the value model to express its internal token-level understanding. We then guide this understanding using auxiliary losses based on entropy and perplexity, encouraging semantic consistency with the actor model despite noisy rewards. Let $P_V(y_t|x)$ denote the token-level prediction distribution from the LM head. We define the auxiliary losses as:

$$L_{\text{ent}} = \sum_{t\in T_{\text{ent}}} H[P_V(y_t|x)], L_{\text{ppl}} = \sum_{t\in T_{\text{ppl}}} -\log P_V(y_t = y_t^*|x) \tag{11}$$

where $H[\cdot]$ is entropy, and T_{ent}, T_{ppl} are dynamically selected token subsets with high entropy and perplexity respectively:

$$T_{\text{ent}} = \{t : H[P_V(y_t|x)] > \hat{T}_{\text{entropy}}\}, \tag{12}$$

$$T_{\text{ppl}} = \{t : -\log P_V(y_t = y_t^*|x) > \hat{T}_{\text{perplexity}}\} \tag{13}$$

where $\hat{T}_{\text{entropy}}$ and $\hat{T}_{\text{perplexity}}$ represent the threshold of entropy and perplexity. The final semantic regularization loss is:

$$L_{\text{semantic}} = \lambda_{\text{ent}} L_{\text{ent}} + \lambda_{\text{ppl}} L_{\text{ppl}} \tag{14}$$

where $\lambda_{\text{ent}}$and λ_{ppl} represent the weight of the loss.

This method provides the value model with stable, semantically meaningful supervision, helping it stay anchored to the input's linguistic structure and resist reward noise. Combined with the information bottleneck from Section 3.2, it reinforces both semantic alignment and robustness.

4 Experiments

4.1 Setup

We evaluate VRPO in two distinct but complementary settings: **rule-based rewards** and **model-based rewards**, both designed to test robustness under noisy supervision.

Rule-based reward setting This scenario mainly **considers the noise in unsupervised learning**. We simulate supervision by performing majority voting over multiple outputs from pretrained models. Two variants are considered: (i) *training-time augmentation* and (ii) *test-time augmentation*, where noisy pseudo-labels from the training/test set supervise RL training. Models are cold-started via supervised fine-tuning and constrained to specific answer formats.

Model-based reward setting We conduct experiments on multi-turn dialogue, using a separate reward model trained on annotated dialogue data. A policy model is optimized via RL under this noisy reward. Performance is measured by interactions with GPT-4o and evaluated with standard dialogue criteria.

Baselines In both settings, we compare VRPO with PPO and GRPO under identical initialization. All models are based on Qwen2.5 (Qwen et al. 2025), Qwen3 (Yang et al. 2025) or Llama3.1 variants, fine-tuned appropriately.

Implementation Details Entropy/perplexity auxiliary loss weight is set to 0.5, activated for 80% of words. Majority voting uses 5 or 32 samples depending on the setting. All models are trained on 8×A100-80GB GPUs.

For further dataset construction, training details, and model initialization, please refer to Appendix.

4.2 Evaluation Metrics

To evaluate the effectiveness of our value model, we consider both task-level performance and value estimation capabilities across different task domains.

Task-Level Metrics For mathematical reasoning and scientific knowledge tasks, we adopt accuracy as the primary metric to measure whether the model produces correct answers. For multi-turn dialogue tasks, we employ a tripartite evaluation framework that captures both task execution and communicative quality: (1) Task Completion Rate (TCR), assessing whether the model successfully fulfills the user-intended objective; (2) Ask Completion Rate (ACR), measuring whether detailed aspects of the user query are adequately addressed; and (3) Goal Completion Rate (GCR), evaluating the fluency, coherence, and appropriateness of the model's responses. This comprehensive evaluation allows us to examine not only whether the model achieves the intended outcomes, but also how effectively it aligns with human communication standards.

Value Model Performance Metrics To directly evaluate the accuracy and robustness of the value model, we adopt the following criteria:

Explained Variance We use the explained variance to measure how much of the empirical return is captured by the predicted value:

$$\text{Explained Variance} = 1 - \frac{\text{Var}[\hat{R} - V]}{\text{Var}[\hat{R}]} \tag{15}$$

where $\hat{R}$ denotes the empirical return and V is the predicted value. A value near 1 indicates high explanatory power, while values near or below zero indicate poor or misleading predictions.

Prediction Error Evaluation To assess the value model's prediction accuracy, we adopt Mean Squared Error (MSE) and Mean Absolute Error (MAE) between the predicted value $V_\theta(s_t)$ and the reference target $V_{\text{target}}(s_t)$:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{T} \sum_{t=0}^{T-1} \left(V_\theta(s_t) - V_{\text{target}}(s_t) \right)^2, \tag{16}$$

$$\mathcal{L}_{\text{MAE}} = \frac{1}{T} \sum_{t=0}^{T-1} \left| V_\theta(s_t) - V_{\text{target}}(s_t) \right| \tag{17}$$

Here, T denotes the trajectory length, and $V_{\text{target}}(s_t)$ is computed via n-step Temporal Difference (TD) with Generalized Advantage Estimation (GAE):

$$V_{\text{target}}(s_t) = \sum_{k=0}^{n-1} \gamma^k r_{t+k} + \gamma^n V(s_{t+n}) \tag{18}$$

If $t + n \geq T$, the bootstrap term $\gamma^n V(s_{t+n})$ is set to zero to account for episode termination. MSE emphasizes large errors and captures prediction stability, while MAE reflects overall distributional deviation. As the rewards are obtained through simulation, **these metrics primarily serve as relative indicators of training quality rather than absolute performance benchmarks**.

Advantage Visualization for Robustness Analysis To qualitatively assess the robustness introduced by value model training, we visualize token-level advantage values across generated sequences. In reinforcement learning-based language generation, the advantage function quantifies the relative benefit of taking action a_t in state s_t, and is computed as:

$$A_t = Q(s_t, a_t) - V(s_t) \tag{19}$$

where $Q(s_t, a_t)$ represents the estimated return of taking action a_t in state s_t, and $V(s_t)$ denotes the baseline value of the current state.

Higher advantage values indicate that the corresponding token exhibits significantly better performance compared to the average policy behavior under the current context. These

tokens function as positive learning signals and reflect directions along which the policy should be reinforced. Therefore, visualizing token-level advantages enables us to analyze which parts of the generation are being prioritized during optimization and how the learned policy aligns with high-reward behaviors.

4.3 Dialogue Task under Model-Based Reward RL Training

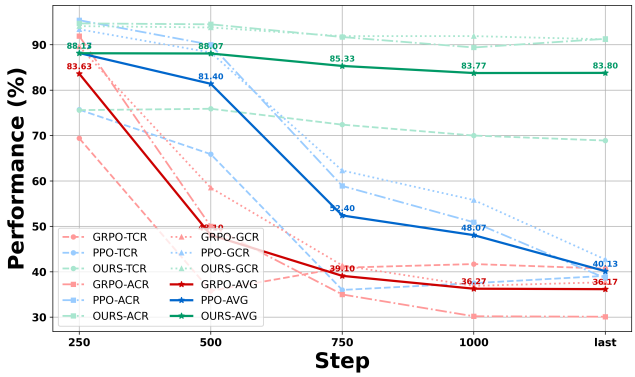

Figure 3: Training performance on dialogue tasks. VRPO (the top lines) improves task completion from 72.1% to 75.9%, avoids collapse under noise, and achieves a final average performance of 83.80%, outperforms GRPO (36.17%) and PPO (40.13%) in stability.

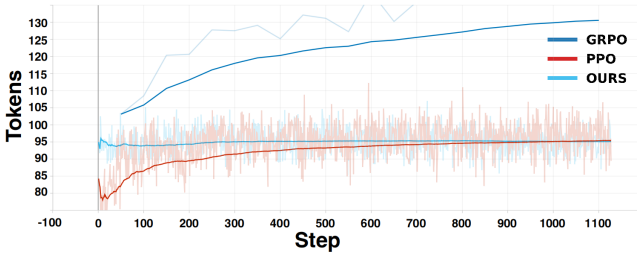

Figure 4: Average response length across training. Our method maintains stable lengths (94 to 95), while PPO and GRPO increase significantly due to length reward hacking.

To further evaluate the effectiveness of our approach in realistic scenarios where supervision is provided by model-based rewards, we conduct experiments on multi-turn dialogue tasks under noisy reward model supervision. Specifically, we utilize the **Honor-Dialogue Dataset** for both reward model training and reinforcement learning. This dataset, which we constructed ourselves, comprises multi-domain task dialogues collected from real-world conversational scenarios. Detailed dataset statistics are provided in the Appendix. Figure 3 illustrates the training dynamics.

VRPO enhances training stability Despite noisy supervision, our method effectively improves dialogue agent performance, surpassing the cold-start baseline and matching PPO's peak performance. In contrast, PPO and GRPO exhibit severe performance collapse under noisy rewards, with GRPO performing the worst. This demonstrates our method's stability and resilience in noisy settings. In particular, the inclusion of a value model is critical: PPO, which retains a value model unlike GRPO, performs significantly better in early training, underscoring the role of the value model in mitigating noise and maintaining learning. For further experimental data results, please refer to Appendix.

Strong Value Model Helps Reduce Reward Hacking A stronger value model helps mitigate reward model hacking by resisting length bias during training. As shown in Figure 4, the performance collapse largely stems from length-based reward exploitation. Our method stabilizes advantage estimation, preventing it from scaling with response length. In contrast, PPO and GRPO exhibit sharp length inflation, revealing greater vulnerability to biased rewards.

4.4 Test-Time Optimization under Rule-Based Rewards

To further assess the effectiveness of our approach in realistic settings where supervision comes from simple rule-based rewards, we conduct a set of experiments covering both mathematical and factual reasoning domains. Specifically, we evaluate models on seven datasets, including four math-focused tasks MATH500 (Hendrycks et al. 2021), AIME24, Minerva-Math (Lewkowycz et al. 2022), AMC23 and three factual knowledge tasks SampleQA (Wei et al. 2024), GPQA , and HLE (Humanity's Last Exam), reporting test-time training results with rule-based feedback. Table 1 presents accuracy after reasoning-enhanced training using different RL methods.

Multi-Domain Improvements via VRPO As shown in Table 1, our method consistently outperforms other baselines across a variety of domains by leveraging a strong value model with semantic-aware filtering. The improvements are evident in both mathematical and knowledge reasoning tasks. For example, accuracy on AIME24 improves from 6.67% to 13.33%, and on HLE from 2.55% to 4.45%. These results highlight the model's ability to extract relevant information from noisy or ambiguous feedback and demonstrate robust generalization across diverse domains.

Toward Meaningful Learning by Value Model under Noisy Supervision Noise is pervasive in both training and test-time data, especially under rule-based feedback. Figure 6 illustrates the error rate in various reward signals. The data noise in the test-time training originates from the majority vote generated by the Qwen2.5-7b-Cold Start model.

Despite noisy feedback, our value model learns effectively. As shown in Figure 5, the explained variance of our method increases steadily during training, indicating closer alignment between predicted and GAE-derived returns. This trend is accompanied by a consistent drop in reward prediction error and reduction in high-variance outliers.

Moreover, our value model helps guide the policy toward more semantically meaningful regions. Figure 2 visualizes

Domain	Math				Factuality	Science	Knowledge	ALL
Method	MATH500	AIME24	Minerva-Math	AMC23	SampleQA	GPQA	HLE	AVG
Cold Start	71.60	6.67	18.75	52.50	1.62	2.53	2.55	22.32
GRPO	73.40	11.67	19.49	56.67	2.47	3.08	3.80	24.37
PPO	73.60	10.00	19.49	55.00	2.43	2.54	3.56	23.80
Ours	**74.40**	**13.33**	**20.22**	**61.67**	**2.61**	**4.17**	**4.45**	**25.84**

Table 1: Accuracy (%) on test-time optimization with rule-based rewards. Our method achieves gains across various datasets and domains.

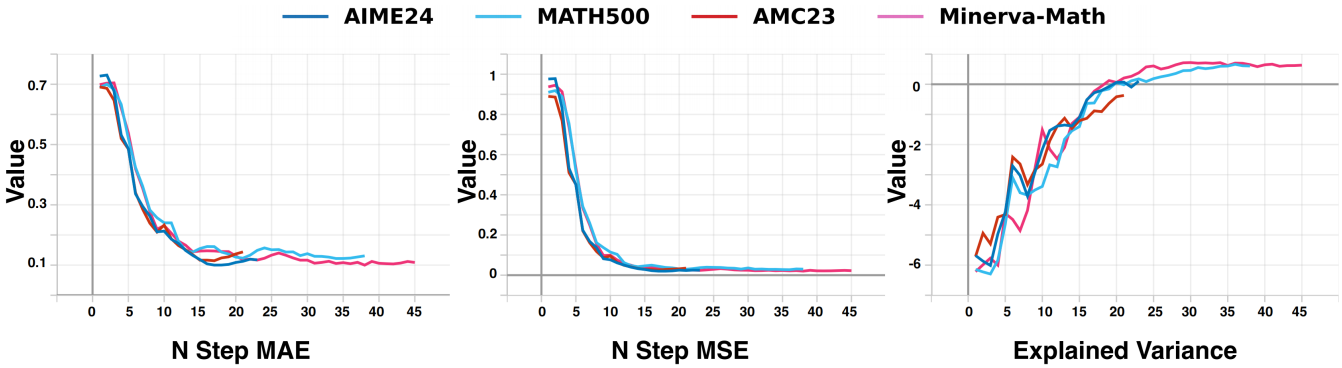

Figure 5: Prediction error of the value model across training steps.

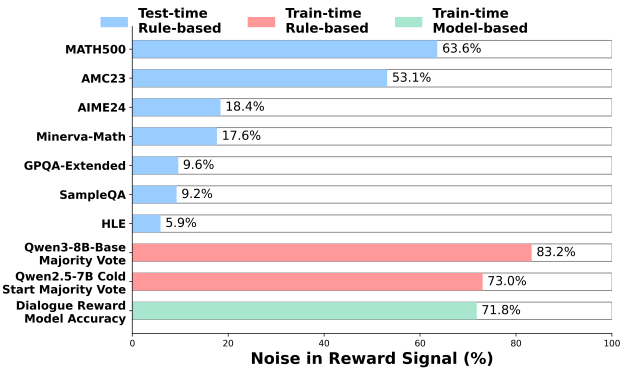

Figure 6: Noise statistics in the various tasks. A significant portion of rewards contains inaccuracies.

the advantage estimation for the same response across different methods. While PPO exhibits dispersed attention across tokens, our model focuses on the key textual information. It not only captures the critical elements of the response but also demonstrates a more appropriate perception and judgment of its answer.

These results demonstrate that even under noisy, rule-based feedback, our value model is able to denoise supervision signals and steer learning toward more meaningful updates, thereby enhancing test-time training effectiveness.

4.5 Train-Time Optimization under Rule-Based Rewards

To evaluate the robustness of the reinforcement experiments when different reinforcement learning methods are trained under the supervision of rule-based rewards, we conducted experiments using the Light-R1 (Wen et al. 2025) dataset for training and the four previous math datasets for testing. Table 2 reports the accuracy of models before and after RL training on both weak and strong base models.

We can observe that:(1) On math tasks with the weaker model (Qwen2.5-7B cold-start), GRPO suffers a significant performance collapse due to the absence of a value model, with accuracy on MATH500 dropping to 50.40%. In contrast, our approach remains robust, benefiting from effective denoising and reward utilization. (2) With a stronger model (Qwen3-8B), the impact of noise diminishes and all methods improve. Nonetheless, our method continues to yield stable gains—achieving 60.55% average accuracy on math datasets, outperforming PPO by 2.15%. The gain is even larger (7.43%) in the weaker model setting, highlighting our method's advantage under more adverse conditions. (3) These results suggest that the benefit of our value model depends on both model capacity and reward noise: while essential in weaker models, it becomes less critical as model strength increases and reward signals become clearer.

4.6 Ablation Experiment and Discussion

Effect of Variational Bottleneck on Noise Filtering As shown in Table 3, we compare different information bottleneck designs across tasks and obtain three takeaways: (1) A single-layer bottleneck delivers the best overall perfor-

Model	MATH500	AIME24	Minerva-Math	AMC23	AVG
Qwen2.5-7B (cold start)					
Base	71.60	6.67	18.75	**52.50**	37.38
GRPO	50.40	0.00	13.60	22.50	21.63
PPO	67.40	13.33	19.12	35.83	33.92
Ours	**72.20**	**23.33**	**19.85**	50.00	**41.35**
Qwen3-8B					
Base	87.40	**41.67**	28.68	75.83	58.40
GRPO	89.20	35.00	28.68	84.17	59.26
PPO	86.40	36.67	30.51	80.00	58.40
Ours	**90.20**	33.33	**31.99**	**86.67**	**60.55**

Table 2: Accuracy (%) on train-time optimization with rule-based rewards. Our method shows the most significant gains under high-noise and low-capacity settings.

Domain	Math				Factuality	Science	Knowledge
Method	**MATH500**	**AIME24**	**Minerva-Math**	**AMC23**	**SampleQA**	**GPQA**	**HLE**
Zero bottleneck	70.20%	**16.67%**	19.49%	50.83%	2.47%	2.17%	3.80%
Two bottleneck	73.20%	**16.67%**	**21.32%**	60.00%	2.45%	3.44%	3.62%
One bottleneck(**Ours**)	**74.40%**	13.33%	20.22%	**61.67%**	**2.61%**	**4.17%**	**4.45%**

Table 3: Performance comparison across domains and datasets using different bottleneck strategies.

mance, with 4.17% and 4.45% accuracy on GPQA and HLE, indicating strong noise-filtering capability. (2) A two-layer bottleneck excels on high-complexity math tasks but underperforms on others, likely due to overfitting to noise. However, for language models like Qwen, which are pretrained on rich mathematical data, deeper value structures may facilitate deeper understanding, yielding benefits in math domains. (3) No bottleneck yields modest gains on low-noise datasets but collapses on noisy ones (e.g., AMC23 drops to 50.83%), showing high susceptibility to noisy or uncertain rewards. Overall, moderate architectural bottlenecks substantially improve the robustness and cross-task generalization of value modeling under noisy RL Training.

Enhanced Semantic Awareness rather than Semantic Understanding As shown in Table 4, we conduct an ablation study to evaluate how different forms of language-awareness affect model robustness under noisy reward supervision. Our results suggest that directly solving tasks via cross-entropy is suboptimal for the value model. Instead, regularizing with entropy and perplexity, focusing on semantic perception, leads to stronger generalization and robustness. Specifically, minimizing these losses improves the value model's ability to identify meaningful signals amid noisy feedback, boosting average accuracy to 42.41%, compared to 39.51% under standard cross-entropy training. This improvement stems from the model's enhanced alignment with linguistic structure, enabling it to act not as a task solver, but as a semantic signal regulator—assessing responses with higher fidelity and filtering unstable patterns. While actor improvements help with comprehension, our results show that the value model is more effective in absorbing reward uncertainty, yielding more stable representations and a 4% gain over actor-centric training on average.

Partial Token Activation may Strengthen Semantic Learning We investigate how partially activating high-uncertainty tokens affects robustness in noisy reward training by selectively applying language-aware losses. As shown in Table 5, activating 80% of high-entropy/perplexity tokens yields the best performance (42.41%), demonstrating that focusing supervision on uncertain regions strengthens the value model's alignment with language semantics. Interestingly, fully activating all tokens (100%) slightly reduces performance (to 41.15%), likely due to the inclusion of noisy or uninformative tokens, which destabilize value estimation. This reveals a trade-off: while broader activation promotes semantic sensitivity, indiscriminate supervision can introduce harmful variance. Partial token activation offers a balanced solution that enhances robustness without compromising core value function stability.

5 Conclusion

In this work, we rethink the often-overlooked role of the value model in the RL framework, especially under noisy or imperfect reward supervision. We propose VRPO, a novel training framework that enhances PPO robustness by incorporating information-bottlenecked architecture and auxiliary supervision guided by entropy and perplexity from a frozen language model, effectively mitigating reward noise propagation. Experiments across reasoning, QA, and dialogue tasks under noisy reward signals demonstrate that VRPO outperforms PPO and GRPO, highlighting the value model's untapped potential in robust RL training.

Method	MATH500	AIME24	Minerva-Math	AMC23	AVG
Actor Model					
Cross-Entropy Loss	73.20%	10.00%	19.12%	50.83%	38.29%
CE with Entropy/Perplexity Filtering	70.40%	13.33%	19.12%	50.83%	38.42%
Entropy + Perplexity Minimization	69.40%	10.00%	18.01%	54.17%	37.90%
Value Model					
Cross-Entropy Loss	73.00%	10.00%	18.38%	56.67%	39.51%
CE with Entropy/Perplexity Filtering	72.20%	10.00%	18.75%	57.50%	39.61%
Entropy + Perplexity Minimization(**Ours**)	**74.40%**	**13.33%**	**20.22%**	**61.67%**	**42.41%**

Table 4: Comparison of different training objectives on math reasoning tasks under test-time reward perturbation. Entropy- and perplexity-based regularization improves value model robustness (+2.9%) over standard cross-entropy, confirming its role as a semantic signal regulator.

Activation	MATH500	AIME24	Minerva-Math	AMC23	AVG
20%	75.00%	6.67%	18.38%	54.17%	38.56%
50%	73.40%	13.33%	19.49%	55.83%	40.51%
80%(**Ours**)	**74.40%**	13.33%	**20.22%**	**61.67%**	**42.41%**
100%	73.80%	**16.67%**	19.12%	55.00%	41.15%

Table 5: Impact of Partial Token Activation on Performance via entropy/perplexity-based loss. Activating 80% of high-uncertainty tokens yields the best accuracy (42.41%).

References

Bellemare, M. G.; Candido, S.; Castro, P. S.; Gong, J.; Machado, M. C.; Moitra, S.; Ponda, S. S.; and Wang, Z. 2020. Autonomous navigation of stratospheric balloons using reinforcement learning. *Nature*, 588: 77 – 82.

Caregnato-Neto, A.; Cavalcante Siebert, L.; Zgonnikov, A.; Omena de Albuquerque Maximo, M. R.; and Junqueira Magalhães Afonso, R. 2024. ARMCHAIR: integrated inverse reinforcement learning and model predictive control for human-robot collaboration. *arXiv e-prints*, arXiv:2402.19128.

Casper, S.; Davies, X.; Shi, C.; Krendl Gilbert, T.; Scheurer, J.; Rando, J.; Freedman, R.; Korbak, T.; Lindner, D.; Freire, P.; Wang, T.; Marks, S.; Segerie, C.-R.; Carroll, M.; Peng, A.; Christoffersen, P.; Damani, M.; Slocum, S.; Anwar, U.; Siththaranjan, A.; Nadeau, M.; Michaud, E. J.; Pfau, J.; Krasheninnikov, D.; Chen, X.; Langosco, L.; Hase, P.; Bıyık, E.; Dragan, A.; Krueger, D.; Sadigh, D.; and Hadfield-Menell, D. 2023. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. *arXiv e-prints*, arXiv:2307.15217.

Chaudhari, S.; Aggarwal, P.; Murahari, V.; Rajpurohit, T.; Kalyan, A.; Narasimhan, K.; Deshpande, A.; and da Silva, B. C. 2024. RLHF Deciphered: A Critical Analysis of Reinforcement Learning from Human Feedback for LLMs. arXiv:2404.08555.

Chen, Y.; Wu, T.; Wang, S.; Feng, X.; Jiang, J.; McAleer, S. M.; Geng, Y.; Dong, H.; Lu, Z.; Zhu, S.-C.; and Yang, Y. 2022. Towards Human-Level Bimanual Dexterous Manipulation with Reinforcement Learning. *arXiv e-prints*, arXiv:2206.08686.

Cheng, J.; Xiong, G.; Dai, X.; Miao, Q.; Lv, Y.; and Wang, F.-Y. 2024. RIME: Robust Preference-based Reinforcement Learning with Noisy Preferences. *arXiv e-prints*, arXiv:2402.17257.

Cheng, R.; Ma, H.; Wang, W.; Wang, Z.; Jia, X.; Qin, S.; Cao, X.; Liu, Y.; and Jia, X. 2025. Inverse Reinforcement Learning with Dynamic Reward Scaling for LLM Alignment. *arXiv e-prints*, arXiv:2503.18991.

Dai, B.; Zhu, C.; and Wipf, D. 2018. Compressing Neural Networks using the Variational Information Bottleneck. *arXiv e-prints*, arXiv:1802.10399.

Gao, L.; Schulman, J.; and Hilton, J. 2022. Scaling Laws for Reward Model Overoptimization. *arXiv e-prints*, arXiv:2210.10760.

Goyal, A.; Islam, R.; Strouse, D.; Ahmed, Z.; Botvinick, M.; Larochelle, H.; Bengio, Y.; and Levine, S. 2019. InfoBot: Transfer and Exploration via the Information Bottleneck. *arXiv e-prints*, arXiv:1901.10902.

Hafner, D.; Lillicrap, T.; Ba, J.; and Norouzi, M. 2019. Dream to Control: Learning Behaviors by Latent Imagination. *arXiv e-prints*, arXiv:1912.01603.

Hendrycks, D.; Burns, C.; Kadavath, S.; Arora, A.; Basart, S.; Tang, E.; Song, D.; and Steinhardt, J. 2021. Measuring Mathematical Problem Solving With the MATH Dataset. arXiv:2103.03874.

Hung, Y.-H.; Hsieh, P.-C.; Mete, A.; and Kumar, P. R. 2023. Value-Biased Maximum Likelihood Estimation for Model-based Reinforcement Learning in Discounted Linear MDPs. arXiv:2310.11515.

Ibarz, B.; Leike, J.; Pohlen, T.; Irving, G.; Legg, S.; and Amodei, D. 2018. Reward learning from human pref-

erences and demonstrations in Atari. *arXiv e-prints*, arXiv:1811.06521.

Laidlaw, C.; Singhal, S.; and Dragan, A. 2025. Correlated Proxies: A New Definition and Improved Mitigation for Reward Hacking. arXiv:2403.03185.

Lewkowycz, A.; Andreassen, A.; Dohan, D.; Dyer, E.; Michalewski, H.; Ramasesh, V.; Slone, A.; Anil, C.; Schlag, I.; Gutman-Solo, T.; Wu, Y.; Neyshabur, B.; Gur-Ari, G.; and Misra, V. 2022. Solving Quantitative Reasoning Problems with Language Models. arXiv:2206.14858.

Liang, X.; Chen, C.; Qiu, S.; Wang, J.; Wu, Y.; Fu, Z.; Shi, Z.; Wu, F.; and Ye, J. 2024. ROPO: Robust Preference Optimization for Large Language Models. *arXiv e-prints*, arXiv:2404.04102.

Martini, K. M.; Abdelaleem, E.; and Nemenman, I. 2024. Deep Variational Multivariate Information Bottleneck. In *APS March Meeting Abstracts*, volume 2024 of *APS Meeting Abstracts*, T28.009.

Miao, Y.; Zhang, S.; Ding, L.; Bao, R.; Zhang, L.; and Tao, D. 2024. InfoRM: Mitigating Reward Hacking in RLHF via Information-Theoretic Reward Modeling. *arXiv e-prints*, arXiv:2402.09345.

Muennighoff, N.; Yang, Z.; Shi, W.; Li, X. L.; Fei-Fei, L.; Hajishirzi, H.; Zettlemoyer, L.; Liang, P.; Candès, E.; and Hashimoto, T. 2025. s1: Simple test-time scaling. arXiv:2501.19393.

Perolat, J.; De Vylder, B.; Hennes, D.; Tarassov, E.; Strub, F.; de Boer, V.; Muller, P.; Connor, J. T.; Burch, N.; Anthony, T.; McAleer, S.; Elie, R.; Cen, S. H.; Wang, Z.; Gruslys, A.; Malysheva, A.; Khan, M.; Ozair, S.; Timbers, F.; Pohlen, T.; Eccles, T.; Rowland, M.; Lanctot, M.; Lespiau, J.-B.; Piot, B.; Omidshafiei, S.; Lockhart, E.; Sifre, L.; Beauguerlange, N.; Munos, R.; Silver, D.; Singh, S.; Hassabis, D.; and Tuyls, K. 2022. Mastering the game of Stratego with model-free multiagent reinforcement learning. *Science*, 378(6623): 990–996.

Poiani, R.; Curti, G.; Metelli, A. M.; and Restelli, M. 2024. Inverse Reinforcement Learning with Sub-optimal Experts. *arXiv e-prints*, arXiv:2401.03857.

Qwen; :; Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; Lin, H.; Yang, J.; Tu, J.; Zhang, J.; Yang, J.; Yang, J.; Zhou, J.; Lin, J.; Dang, K.; Lu, K.; Bao, K.; Yang, K.; Yu, L.; Li, M.; Xue, M.; Zhang, P.; Zhu, Q.; Men, R.; Lin, R.; Li, T.; Tang, T.; Xia, T.; Ren, X.; Ren, X.; Fan, Y.; Su, Y.; Zhang, Y.; Wan, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; and Qiu, Z. 2025. Qwen2.5 Technical Report. arXiv:2412.15115.

Schulman, J.; Moritz, P.; Levine, S.; Jordan, M.; and Abbeel, P. 2018. High-Dimensional Continuous Control Using Generalized Advantage Estimation. arXiv:1506.02438.

Wang, B.; Zheng, R.; Chen, L.; Liu, Y.; Dou, S.; Huang, C.; Shen, W.; Jin, S.; Zhou, E.; Shi, C.; Gao, S.; Xu, N.; Zhou, Y.; Fan, X.; Xi, Z.; Zhao, J.; Wang, X.; Ji, T.; Yan, H.; Shen, L.; Chen, Z.; Gui, T.; Zhang, Q.; Qiu, X.; Huang, X.; Wu, Z.; and Jiang, Y.-G. 2024a. Secrets of RLHF in Large Language Models Part II: Reward Modeling. *arXiv e-prints*, arXiv:2401.06080.

Wang, C.; Gan, Y.; Huo, Y.; Mu, Y.; Yang, M.; He, Q.; Xiao, T.; Zhang, C.; Liu, T.; Du, Q.; Yang, D.; and Zhu, J. 2024b. RoVRM: A Robust Visual Reward Model Optimized via Auxiliary Textual Preference Data. *arXiv e-prints*, arXiv:2408.12109.

Wang, X.; Dou, S.; Xiong, L.; Zou, Y.; Zhang, Q.; Gui, T.; Qiao, L.; Cheng, Z.; and Huang, X. 2022. MINER: Improving Out-of-Vocabulary Named Entity Recognition from an Information Theoretic Perspective. *arXiv e-prints*, arXiv:2204.04391.

Wei, J.; Karina, N.; Chung, H. W.; Jiao, Y. J.; Papay, S.; Glaese, A.; Schulman, J.; and Fedus, W. 2024. Measuring short-form factuality in large language models. arXiv:2411.04368.

Wen, L.; Cai, Y.; Xiao, F.; He, X.; An, Q.; Duan, Z.; Du, Y.; Liu, J.; Tang, L.; Lv, X.; Zou, H.; Deng, Y.; Jia, S.; and Zhang, X. 2025. Light-R1: Curriculum SFT, DPO and RL for Long COT from Scratch and Beyond. arXiv:2503.10460.

Wu, J.; Xie, Y.; Yang, Z.; Wu, J.; Chen, J.; Gao, J.; Ding, B.; Wang, X.; and He, X. 2024. Towards Robust Alignment of Language Models: Distributionally Robustifying Direct Preference Optimization. *arXiv e-prints*, arXiv:2407.07880.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; Zheng, C.; Liu, D.; Zhou, F.; Huang, F.; Hu, F.; Ge, H.; Wei, H.; Lin, H.; Tang, J.; Yang, J.; Tu, J.; Zhang, J.; Yang, J.; Yang, J.; Zhou, J.; Zhou, J.; Lin, J.; Dang, K.; Bao, K.; Yang, K.; Yu, L.; Deng, L.; Li, M.; Xue, M.; Li, M.; Zhang, P.; Wang, P.; Zhu, Q.; Men, R.; Gao, R.; Liu, S.; Luo, S.; Li, T.; Tang, T.; Yin, W.; Ren, X.; Wang, X.; Zhang, X.; Ren, X.; Fan, Y.; Su, Y.; Zhang, Y.; Zhang, Y.; Wan, Y.; Liu, Y.; Wang, Z.; Cui, Z.; Zhang, Z.; Zhou, Z.; and Qiu, Z. 2025. Qwen3 Technical Report. arXiv:2505.09388.

Yue, T.; Cheng, J.; Guo, L.; Dai, X.; Zhao, Z.; He, X.; Xiong, G.; Lv, Y.; and Liu, J. 2024. SC-Tune: Unleashing Self-Consistent Referential Comprehension in Large Vision Language Models. *arXiv e-prints*, arXiv:2403.13263.

Zhang, J.; Jing, W.; Zhang, Z.; Xi, Z.; Dou, S.; Weng, R.; Li, J.; Wang, J.; Chai, M.; Hong, S.; Gui, T.; and Zhang, Q. 2025. Two Minds Better Than One: Collaborative Reward Modeling for LLM Alignment. *arXiv e-prints*, arXiv:2505.10597.

Zhou, M.; Luo, J.; Villella, J.; Yang, Y.; Rusu, D.; Miao, J.; Zhang, W.; Alban, M.; Fadakar, I.; Chen, Z.; Chongxi Huang, A.; Wen, Y.; Hassanzadeh, K.; Graves, D.; Chen, D.; Zhu, Z.; Nguyen, N.; Elsayed, M.; Shao, K.; Ahilan, S.; Zhang, B.; Wu, J.; Fu, Z.; Rezaee, K.; Yadmellat, P.; Rohani, M.; Perez Nieves, N.; Ni, Y.; Banijamali, S.; Cowen Rivers, A.; Tian, Z.; Palenicek, D.; Ammar, H. b.; Zhang, H.; Liu, W.; Hao, J.; and Wang, J. 2020. SMARTS: Scalable Multi-Agent Reinforcement Learning Training School for Autonomous Driving. *arXiv e-prints*, arXiv:2010.09776.

A Additional Details for VRPO

A.1 Pseudocode

The full algorithm of VRPO is detailed in Algorithm 1.

Algorithm 1: VRPO training process

Require: Dataset $\mathcal{D} = \{(x_t, a_t, r_t)\}_{t=1}^{T}$, policy π_θ, decoder model q_ψ, encoder model f_ϕ, bottleneck prior $r(z)$; coefficients $\beta, \lambda_{\text{ent}}, \lambda_{\text{ppl}}$

Ensure: Trained actor π_θ and value model f_ϕ, q_ψ

for each VRPO iteration **do**
  Sample trajectories using current policy π_θ
  Compute rewards $\{r_t\}$ and actions $\{a_t\}$
  Compute policy ratio: $\rho_t = \frac{\pi_\theta(a_t|x_t)}{\pi_{\theta_{\text{old}}}(a_t|x_t)}$
  for each timestep t **do**
    Encode latent: sample $\epsilon \sim \mathcal{N}(0, I)$, compute $z_t = f_\phi^\mu(x_t) + f_\phi^\sigma(x_t) \cdot \epsilon$
    Predict value: $\hat{V}_t = q_\psi(z_t)$
  end for
  Compute GAE advantage: $A_t = \text{GAE}(r_t, \hat{V}_t, \hat{V}_{t+1})$
  Compute returns: $R_t = A_t + \hat{V}_t$
  Value Model Update
  // Value Loss with IB Structure
  Compute value loss: $\mathcal{L}_{\text{MSE}} = \frac{1}{T}\sum_t (q_\psi(f\phi(x_t)) - R_t)^2$
  Compute KL loss: $\mathcal{L}_{\text{KL}} = \frac{1}{T}\sum_t \text{KL}(p_\phi(z_t|x_t) \| r(z))$
  // Semantic Alignment via Frozen LM Head
  for each timestep t **do**
    Get token distribution $P_V(y_t|x_t)$ from frozen LM head
    Identify high-uncertainty tokens:
      $T_{\text{ent}} = \{t : H[P_V(y_t|x_t)] > \hat{T}_{\text{entropy}}\}$
      $T_{\text{ppl}} = \{t : -\log P_V(y_t = y_t^*|x_t) > \hat{T}_{\text{perplexity}}\}$
    Compute entropy loss: $L_{\text{ent}} = \sum_{t \in T_{\text{ent}}} H[P_V(y_t|x_t)]$
    Compute perplexity loss: $L_{\text{ppl}} = \sum_{t \in T_{\text{ppl}}} -\log P_V(y_t = y_t^*|x_t)$
  end for
  Compute semantic loss: $\mathcal{L}_{\text{sem}} = \lambda_{\text{ent}} \cdot L_{\text{ent}} + \lambda_{\text{ppl}} \cdot L_{\text{ppl}}$
  Combine value loss:

$$\mathcal{L}_{\text{value}} = \mathcal{L}_{\text{MSE}} + \beta \cdot \mathcal{L}_{\text{KL}} + \mathcal{L}_{\text{sem}}$$

  Update value model $\{\phi, \psi\}$ using $\nabla \mathcal{L}_{\text{value}}$
  Actor Model Update
  Compute clipped PPO objective:

$$\mathcal{L}_{\text{PPO}} = \mathbf{E}_t \left[\min\left(\rho_t A_t, \text{clip}(\rho_t, 1-\epsilon, 1+\epsilon) A_t\right)\right]$$

  Update actor model π_θ, using $\nabla \mathcal{L}_{\text{PPO}}$
end for

B Additional Experimental Details

B.1 Setup

Dataset Construction

Rule-based Setting. For **test-time augmentation**, we use four math benchmarks: AIME 2024, AMC 2023, Minerva-Math, and MATH-500, as well as three QA datasets: GPQA-Extended, SimpleQA, and HLE (Humanity's Last Exam).

For **training-time augmentation**, pseudo-labels are generated using 39,000 samples from the Light-R1 dataset. After 5-round majority voting, samples with at least 3 identical responses are retained, resulting in 31,209 samples for RL training. Evaluation uses math benchmarks with no overlap to the training set.

Model-based Setting. A real-world dialogue dataset (Honor-Dialogue Dataset) is used. The reward model is trained on 36,000 labeled samples, with 3,000 for validation. The policy model is fine-tuned and trained on 80,000 additional conversations, with 8,000 for evaluation.This dataset contains various dialogue tasks in real-life scenarios. It truly reflects the actual conditions of real situations. Such multi-category real-scenario dialogue task data is not available in other datasets.

Baseline Initialization

Rule-based Setting. **Test-time augmentation:** Qwen2.5-7B-Base (math task) and Qwen2.5-7B-Instruct (QA task), fine-tuned on OpenR1-Math and S1.1(Muennighoff et al. 2025), respectively. **Training-time augmentation:** Qwen2.5-7B-Base-ColdStart , Qwen3-8B-Base and Llama-3.1-8B-Instruct are used as baselines.

Model-based Setting. Both the policy and reward models are initialized from Qwen3-8B-Base, fine-tuned on the Honor-Dialogue dataset.

Training Configuration

Training parameters. Majority voting: 32 samples for test-time, 5 for training-time. Auxiliary losses (entropy and perplexity): weight = 0.5, applied on 80% token positions. Experiments are run on 8×NVIDIA A100 80GB GPUs. All other parameters are set properly.

B.2 Additional experimental details in the Dialogue Task

To further analyze the behavior of different methods in the dialogue task, Table 6 presents the full training trajectory across various steps. A notable observation is that both PPO and GRPO experience severe performance degradation during training, likely due to reward over-optimization or instability, resulting in significantly lower final scores than both the initial model and our method. Specifically, their final average scores drop to 40.13% and 36.17%, respectively.

In contrast, our proposed method (VRPO) maintains high robustness throughout training. While the final performance slightly decreases compared to the cold-start model (from 85.87% to 83.80%), our method avoids collapse and achieves high task performance in early training stages. For example, VRPO reaches 75.90% task completion rate (TCR) at step 500 and reaches 88.13% overall average (AVG) score at step 250, outperforming all baselines at these stages.

These results suggest that VRPO effectively preserves model capabilities under heavy noisy supervision and raises the lower bound of post-training performance. This stabilizing effect makes it significantly less prone to reward hacking or collapse compared to standard PPO approaches.

Step	TCR	ACR	GCR	AVG
Cold Start				
0	72.10%	94.30%	91.20%	85.87%
GRPO				
250	69.40%	91.90%	89.60%	83.63%
500	35.80%	50.00%	58.50%	48.10%
750	40.90%	35.00%	41.40%	39.10%
1000	41.70%	30.20%	36.90%	36.27%
Last	40.70%	30.10%	37.70%	36.17%
PPO				
250	75.70%	**95.40%**	93.40%	**88.17%**
500	65.90%	90.00%	88.30%	81.40%
750	36.00%	58.90%	62.30%	52.40%
1000	37.50%	50.90%	55.80%	48.07%
Last	39.10%	38.60%	42.70%	40.13%
OURS				
250	75.60%	94.70%	**94.10%**	88.13%
500	**75.90%**	94.50%	93.80%	88.07%
750	72.40%	91.70%	91.90%	85.33%
1000	70.00%	89.40%	91.90%	83.77%
Last	68.90%	91.30%	91.20%	83.80%

Table 6: Trajectory-level generalization performance across different methods and training steps. Our method maintains high performance even in last stages, beyond GRPO and PPO.

B.3 Additional experimental details under rule-based reward supervision

Tables 7 and 8 present the performance of the LLaMA 3.1-8B-Instruct model on four mathematical reasoning benchmarks under rule-based reward supervision, after unsupervised reinforcement learning. Table 7 shows results for test-time optimization, while Table 8 reports training-time optimization.

We observe that the LLaMA model performs less strongly on mathematical tasks compared to the Qwen model. Notably, in the test-time optimization setting, only our method improves upon the base model, while PPO and GRPO show performance degradation, highlighting our method's robustness under noisy supervision.

In the training-time optimization setting, due to limited model capacity and the prevalence of noisy samples (overall data accuracy is only 42.33%), all methods experience performance drops. However, our method shows the least decline and achieves improvements on Minerva-Math and AMC23, demonstrating stronger resilience in noisy training environments.

B.4 Case Study

The dialogue task experiments based on the Honor-Dialogue Dataset serve as a key part of our evaluation. This dataset is constructed by us and, to the best of our knowledge, is the first real-world, multi-domain dialogue dataset of its kind. Figure 7 shows an example from our dialogue experiments.

Method	MATH 500	AIME 24	Minerva Math	AMC 23	AVG
Base	**49.80%**	**10.00%**	17.28%	24.17%	25.31%
GRPO	49.20%	6.67%	16.18%	**27.50%**	24.89%
PPO	46.20%	6.67%	**18.01%**	25.00%	23.97%
Ours	49.40%	**10.00%**	17.65%	**27.50%**	**26.14%**

Table 7: Test-time optimization results on mathematical reasoning benchmarks using LLaMA 3.1-8B-Instruct under noisy rule-based reward supervision. Our method is the only one that improves upon the base model, demonstrating superior robustness under noisy feedback.

Method	MATH 500	AIME 24	Minerva Math	AMC 23	AVG
Base	**49.80%**	**10.00%**	17.28%	24.17%	**25.31%**
GRPO	42.40%	3.33%	17.28%	**32.50%**	23.88%
PPO	41.00%	6.67%	17.65%	25.83%	22.79%
Ours	41.40%	**10.00%**	**18.01%**	30.00%	24.85%

Table 8: Training-time optimization results under noisy reward conditions. Due to weak base performance and high noise, all methods show a performance drop, but ours degrades the least and even improves results on Minerva-Math and AMC23.

Each model response carries various state signals, and the multi-turn dialogue continues until an explicit termination state is reached. For evaluation, we use **GPT-4o** as the caller, while the model plays the role of the receiver, engaging in multi-turn interactions. The overall performance is assessed based on the outcomes of these multi-turn dialogues.

B.5 Additional Visualization

Figures 9, 10, 11, and 12 present a comparison of advantage estimations between PPO and our method (VRPO) on the same question (Figure 8) outputs. This is also an example demonstrating the effectiveness of our method compared to the PPO method.

Our method demonstrates a clear ability to capture key semantic components. In the correct answer case (Figure 9), it effectively identifies meaningful concepts such as "massive" and "initial mass M", while also showing a higher advantage estimate for the correct answer (Figure 9) compared to the incorrect one (Figure 10). The model not only captures answer structures like box-based outputs but also makes more accurate judgments based on correctness.

In contrast, PPO fails to capture such distinctions. As seen in Figure 11, the advantage estimation for the correct answer collapses, with later tokens receiving increasing advantage values. This pattern can lead to length hacking, where the model learns to generate longer outputs regardless of semantic quality. Furthermore, in the incorrect answer case (Figure 12), PPO assigns high advantages to both relevant and irrelevant content, indicating a lack of focus on key semantic cues and ultimately degrading training performance.

System Prompt

Role: You are the AI call assistant of Honor Inc., capable of answering calls clearly and politely on behalf of the user and conversing with the caller. Based on the user's customized dialogue goals, you should first determine which goal applies to the conversation and respond accordingly. You do not know any personal information about the user beyond the task description and must inform the caller that any information they provide will be passed on to the user.
Dialogue Goals (Customized):

- **0: General Domain** – Used if the call doesn't match any specific domain. Be polite and professional. Try to answer or guide the caller and clarify their intent.
- **1: E-commerce/Retail** – If caller mentions down jacket performance, mention fast delivery, good warmth at -5°C, ask for cleaning tips, other styles, hotline, or exchange policy, inquire about washing temperature.
- **2: Technical Support** – If license issues arise for WX3510/WX5540, answer related questions, confirm email and phone, send manuals, ask for model and associated phone.
- **3: Financial Services – Pension** – If retirement certification reminder, say not remembered, will handle via WeChat, thank them, ask about past communication and WeChat face verification method.
- **4: Life Services** – If laundry fee update, respond with satisfaction, hope for discount, mention night service needs, ask about pricing explanation, training quality, and monthly discounts.
- **5: Financial Services – Trade Notice** – If transaction confirmation, say user unavailable. If medium-risk, ask for analyst opinion, arrange follow-up next day, inquire about transaction fees.

Caller History and New Input

Conversation History: *(Empty)*
Caller New Input: Hello, this is the Forex Trading Customer Service Center. We are confirming that the currency pair you traded is USD/SGD, and the rate information is ready.

Model Response

Target Domain: *5: Financial Services – Trade Notice*
Dialogue State: *Ongoing*
Response: Hello, this call is currently being handled by the smart assistant. The user is temporarily unavailable. If this is a medium-risk transaction, could we get advice from a market analyst? Please contact us again tomorrow for confirmation.

Figure 7: A multi-turn example from the Honor-Dialogue dataset. The dataset features realistic, task-oriented, multi-domain conversations, in which each model response includes structured states. This example represents the situation of the first conversation.

Physics Fission Problem

Problem: An atomic nucleus of mass M is at rest. A spontaneous fission occurs in which the nucleus splits into two fragments (and nothing else), such that one fragment is 2 times more massive than the other (in terms of their rest-masses). The sum of rest-masses of the two fragments is 99% of the initial mass M. If the rest-mass energy of the initial nucleus (before fission) was 300 GeV, what is the kinetic energy of the more massive fragment?
Assumptions: Velocities of the fragments are significantly smaller than the speed of light. Ignore the electrons.

Figure 8: Example problem from the GPQA dataset

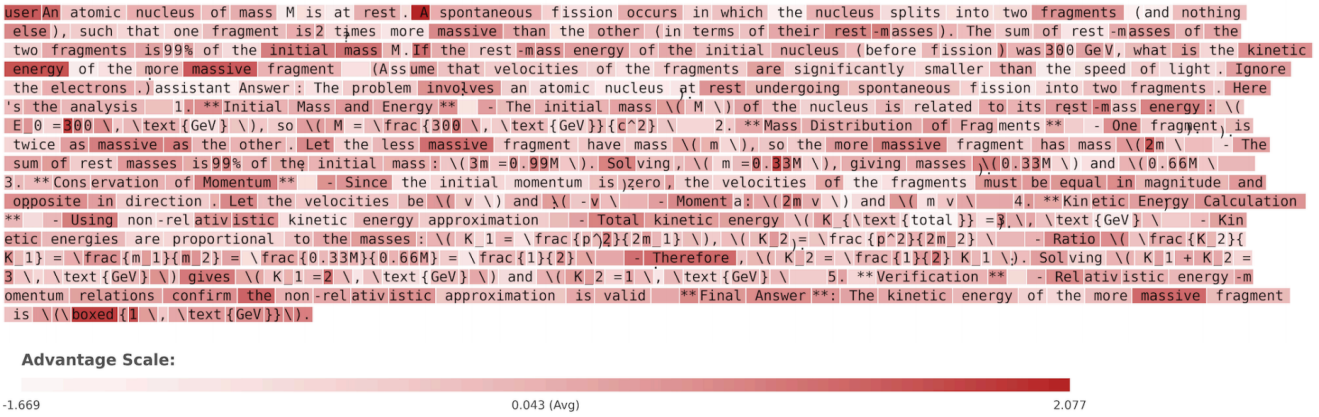

Figure 9: Our method on correct answer. Key information such as ”initial mass M” is successfully captured.

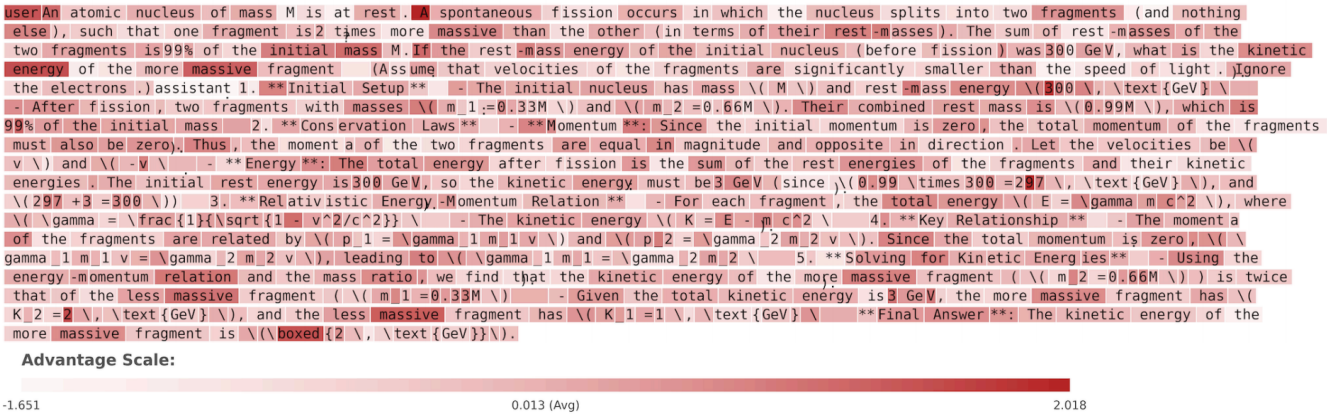

Figure 10: Our method on incorrect answer. The wrong prediction is appropriately assigned low advantage.

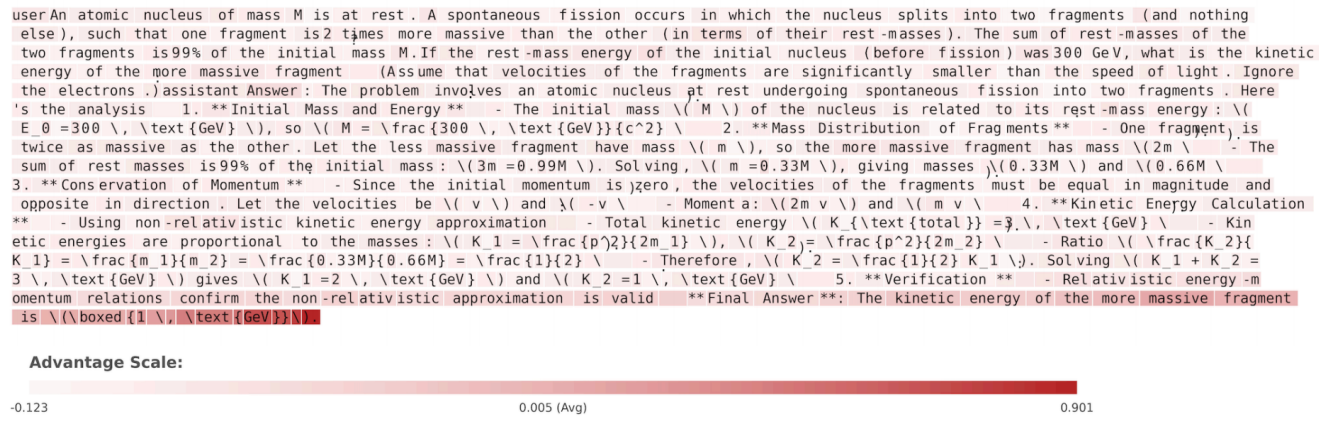

Figure 11: PPO on correct answer. The model fails to capture the semantic core and assigns low or inconsistent advantage.

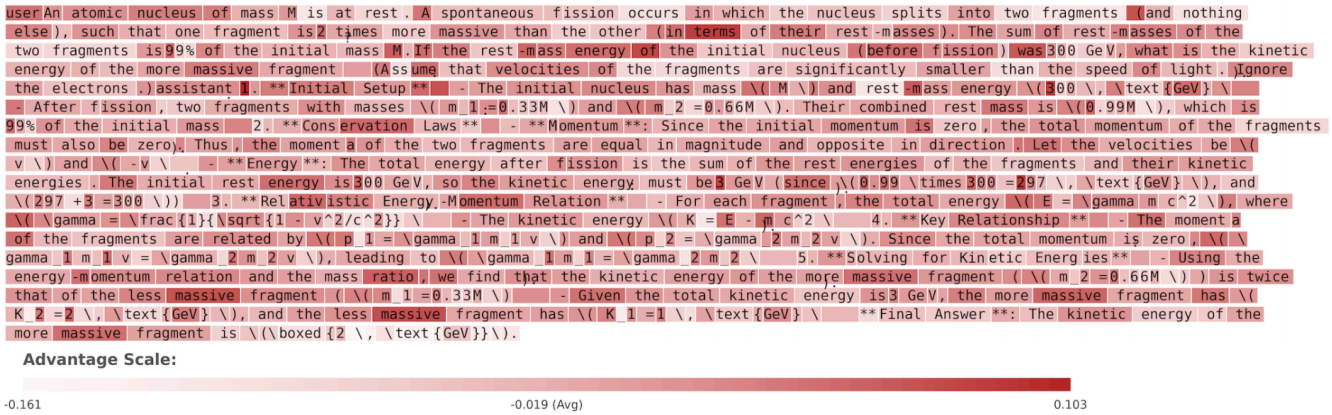

Figure 12: PPO on incorrect answer. High advantage is mistakenly assigned to both key and irrelevant content, weakening semantic discrimination.