Preventing Process Reward Model Hacking When Training Large Language Models on Verifiable Rewards

Extended Abstract

Grant C. Forbes
North Carolina State University
Raleigh, United States
gforbes@alumni.ncsu.edu

Leonardo Villalobos-Arias
North Carolina State University
Raleigh, United States
lvillal@ncsu.edu

Jianxun Wang
North Carolina State University
Raleigh, United States
jwang75@ncsu.edu

Arnav Jhala
North Carolina State University
Raleigh, United States
ahjhala@ncsu.edu

David L. Roberts
North Carolina State University
Raleigh, United States
dlrober4@ncsu.edu

ABSTRACT

Alignment of Large Language Models to human preferences is an active and important field of study. Recent work in Reinforcement Learning with Verifiable Rewards (RLVR) aims to bypass the need for costly and imprecise human preference reward data by training LLMs specifically in domains wherein a simple, known solution exists. As the RLVR signal is sparse, however, it is often supplemented with a Process Reward Model (PRM) reward, which provides a dense reward for each token, or step in the chain of thought, for an agent to learn from. Using the VersaPRM extension to the MMLU-Pro dataset, we demonstrate that PRMs are susceptible to reward hacking behavior, wherein the model is incentivized to produce particularly long, plausible-seeming chains of thought that do not result in the correct response. We also develop a theoretical framework and a suite of methods for preventing this reward hacking while still utilizing PRMs effectively, based on recent work in potential-based and optimality-preserving reward shaping. We both prove theoretically and demonstrate practically that these methods prevent PRMs from altering the optimal policy, and thus from being optimized at the expense of the RLVR signal.

KEYWORDS

Large Language Models; Alignment; Reward Hacking; Reinforcement Learning; Reward Shaping

ACM Reference Format:
Grant C. Forbes, Leonardo Villalobos-Arias, Jianxun Wang, Arnav Jhala, and David L. Roberts. 2026. Preventing Process Reward Model Hacking When Training Large Language Models on Verifiable Rewards: Extended Abstract. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/BWFN6920

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/BWFN6920

1 INTRODUCTION AND BACKGROUND

When training Large Language Models (LLMs) on some combination of a sparse, verifiable reward and a dense, heuristic reward, there is an acute risk of the LLM optimizing for the dense reward at the expense of the more robustly accurate one. Leveraging insights from the fields of intrinsic motivation and optimality-preserving reward shaping, we develop two simple methods for minimizing this risk. We demonstrate, both in theory and practice, that both methods prevent reward hacking in this domain.

Specifically, we extend the theoretical scope of both Generalized Reward Matching (GRM)[4] and Action-Dependent Optimality-Preserving Shaping (ADOPS)[5, 6] to be straightforwardly applicable to training LLMs with Process Reward Models (PRMs)[9] and verifiable rewards (RLVR)[8] concurrently, and prove that these methods' theoretical guarantees still hold in this domain. We also demonstrate that not only are incentives for reward hacking prevalent in SOTA PRM datasets and methods, but that our extensions to optimality-preserving reward shaping effectively mitigate these incentives.

Reward shaping modifies some original reward R by replacing it with $R' = R + F$, where F is the "shaping reward."[1] This can help agents to learn otherwise-sparse problems; however, it can also create incentives for reward hacking, wherein the agent optimizes for F at the expense of R [12]. Recent extensions of Potential-Based Reward Shaping [2, 3, 11] address this problem by providing "plug-and-play" methods for converting fairly general reward shaping terms to a form which provably doesn't alter the optimal policy set of the underlying environment, given a set of fairly general assumptions. We present a novel extension of two of these optimality-preseving methods, Generalized Reward Matching (GRM)[4] and Action-Dependent Optimality-Preserving Shaping (ADOPS)[5, 6], to specifically prevent reward hacking when training LLMs on verifiable rewards with process reward models.

Recent work in RLVR has focused on training LLMs in domains where there exists some "verifiable" reward signal, such as coding or math problems [8]. RLVR objectives, however, are often sparse: PRMs have thus become a common and successful method for supplementing an RLVR objective [10, 15, 16].

[1] In this paper, we will generally refer to R and F as the "intrinsic" and "extrinsic" rewards, respectively.

We explicitly draw on the structural similarity between PRMs and traditional reward shaping to prevent PRM-induced reward hacking. Some prior work relating both of these fields exists, [1, 13], including Gao et al. [7] who discuss the possibility of PRM reward hacking, but ours is the first work to transfer PBRS-based theoretical guarantees to the domain of verifiable rewards with LLMs.

2 PRESERVING OPTIMALITY IN RLVR

LLM training differs from more standard RL environments in multiple ways, and we leverage these dynamic differences to implement more effective and efficient versions of both GRM and ADOPS.

Our new formulation of GRM drops one of the future-agnosticity conditions of Forbes et al. [4], thus becoming more general, easier to implement, and more applicable to LLMs, while still preserving optimality. Assuming that the discount $\gamma = 1$, we define $\bar{F_E}$ to be the arithmetic mean of collected PRM rewards within some episode E. We then find a valid GRM matching function which leads to the shaping reward

$$F_t'^{\text{GRM}} = F_t - \bar{F_E}. \tag{1}$$

This is a particularly natural and easy-to-compute GRM formulation, one that has no risk of of the exponential explosions noted by Forbes et al. [5], and preserves the optimal policy.

We can similarly modify and improve ADOPS for the LLM RLVR domain. We develop an ADOPS extension that foregoes the need to access a critic network, and instead relies on comparing intrinsic and extrinsic *returns* within an episode. This is possible here due to the "dynamics" of the LLM environment while executing a CoT response: each "state" (the current CoT/answer, as well as the prior prompt/CoT history) depends solely and deterministically on the "action" taken by the agent (a particular CoT line). This determinism, combined with the fact that we have access to fully completed trajectories, allows us to simplify ADOPS substantially. We define $U_I^*(s) = \max(U_I^{\tau\sim\pi^*}(s), 0)$ to be the maximum PRM reward obtained by the agent from the state s. With this definition, the new ADOPS shaping reward is $F' = F + F_2$, where

$$F_2(s) = \begin{cases} \min(0, 1 + U_I^*(s) - U_I^{\tau\sim\pi}(s') - F(s) - \epsilon) & \text{if} \quad U_E^{\tau\sim\pi} < 1, \\ \max(0, U_I^*(s) - U_I^{\tau\sim\pi}(s') - F(s)) & \text{if} \quad U_E^{\tau\sim\pi} = 1. \end{cases} \tag{2}$$

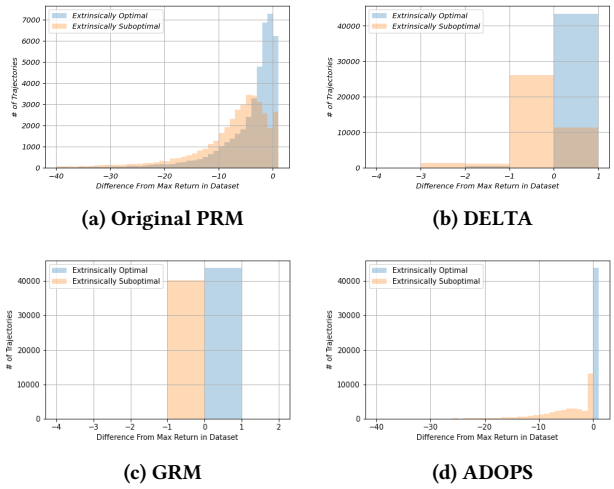
This form of ADOPS provably prevents reward hacking.

In the next section, we analyze both the susceptibility of PRM models to reward hacking and the ability of these newly tailored methods to prevent this hacking.

3 EMPIRICAL ANALYSIS OF PRM HACKING

The VersaPRM dataset, developed by Zeng et al. [16], contains approximately 84,000 trajectories evaluated across 5,750 unique prompts from the MMLU-Pro dataset developed by Wang et al. [14]. Each trajectory contains both a CoT and a final response to the prompt, with an indication of whether the response is correct. Each CoT line is also accompanied by an intrinsic reward label, either -1 for a poor CoT line or 1 for a good one.

To determine if the VersaPRM dataset contains incentives for reward hacking, we group trajectories by shared prompt and found the highest-return trajectory for each prompt (intrinsic + extrinsic).



(a) Original PRM (b) DELTA

(c) GRM (d) ADOPS

Figure 1: The difference from the max of trajectory returns sharing the same prompt included in the VersaPRM dataset with various PRM correction methods. Extrinsically optimal trajectories with a difference < 0 represent a change to the optimal policy, as do extrinsically suboptimal trajectories with a difference of 0.

Reward	TP %	TN %	FP %	FN %
Original	7.4	44.6	3.2	44.8
Delta	51.6	34.2	13.5	0.6
GRM	**52.2**	**47.8**	**0.0**	**0.0**
ADOPS	**52.2**	**47.8**	**0.0**	**0.0**

Table 1: Results of optimalty preservation in the VersaPRM dataset. GRM and ADOPS both fully preserve optimality, while the original reward and the DELTA method [7] do not.

Then, for each trajectory sharing a prompt with the highest trajectory, we calculated the difference between its return and the highest positive-return trajectory. We plot the resulting histogram for the original reward, as well as that modified by DELTA[7], GRM[4] and ADOPS[5], in Figure 1. This plot shows definitively that incentives for reward hacking exist within this dataset: every extrinsically optimal trajectory with a difference less than 0 represents an optimal policy with suboptimal shaped return, and every extrinsically suboptimal trajectory with a difference equal to 0 represents a suboptimal policy with optimal shaped return. Figure 1 also shows that, while DELTA does not fully mitigate this hacking, our adaptations of GRM and ADOPS do. These results are confirmed by the FP and FN rates in Table 1.

Thus, it is important to use GRM, ADOPS, or some similar optimality-preserving method when deploying PRMs in an RLVR setting, lest the noisy dense signal be hacked at the expense of the verifiable rewards whose optimization is our ultimate goal.

REFERENCES

[1] Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, et al. 2024. Do not think that much for 2+ 3=? on the overthinking of o1-like llms. *arXiv preprint arXiv:2412.21187* (2024).

[2] Grant C. Forbes and David L. Roberts. 2024. Potential-Based Reward Shaping For Intrinsic Motivation (Student Abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI.

[3] Grant C Forbes, Leonardo Villalobos-Arias, Nitish Gupta, Colin M Potts, Arnav Jhala, and David L Roberts. 2024. Potential-Based Reward Shaping For Intrinsic Motivation. In *23rd International Conference on Autonomous Agents and Multiagent Systems*. ACM, 589–598.

[4] Grant C Forbes, Leonardo Villalobos-Arias, Jianxun Wang, Arnav Jhala, and David L Roberts. 2024. Potential-Based Intrinsic Motivation: Preserving Optimality With Complex, Non-Markovian Shaping Rewards. *arxiv preprint* (2024).

[5] Grant Collier Forbes, Jianxun Wang, Leonardo Villalobos-Arias, Arnav Jhala, and David Roberts. 2025. Action-Dependent Optimality-Preserving Reward Shaping. In *Forty-second International Conference on Machine Learning*.

[6] Grant C. Forbes, Jianxun Wang, Leonardo Villalobos-Arias, Arnav Jhala, and David I. Roberts. 2025. Action-Dependent Optimality-Preserving Reward Shaping. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems* (Detroit, MI, USA) *(AAMAS '25)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 2523–2525.

[7] Jiaxuan Gao, Shusheng Xu, Wenjie Ye, Weilin Liu, Chuyi He, Wei Fu, Zhiyu Mei, Guangju Wang, and Yi Wu. 2024. On designing effective rl reward at training time for llm reasoning. *arXiv preprint arXiv:2410.15115* (2024).

[8] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948* (2025).

[9] Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

[10] Liangchen Luo, Yinxiao Liu, Rosanne Liu, Samrat Phatale, Meiqi Guo, Harsh Lara, Yunxuan Li, Lei Shu, Yun Zhu, Lei Meng, et al. 2024. Improve mathematical reasoning in language models by automated process supervision. *arXiv preprint arXiv:2406.06592* (2024).

[11] Andrew Y Ng, Daishi Harada, and Stuart Russell. 1999. Policy invariance under reward transformations: Theory and application to reward shaping. In *ICML*, Vol. 99. 278–287.

[12] Jette Randløv and Preben Alstrøm. 1998. Learning to Drive a Bicycle Using Reinforcement Learning and Shaping.. In *ICML*, Vol. 98. Citeseer, 463–471.

[13] Yanming Wan, Jiaxing Wu, Marwa Abdulhai, Lior Shani, and Natasha Jaques. 2025. Enhancing Personalized Multi-Turn Dialogue with Curiosity Reward. *arXiv preprint arXiv:2504.03206* (2025).

[14] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems* 37 (2024), 95266–95290.

[15] Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A Smith, Mari Ostendorf, and Hannaneh Hajishirzi. 2023. Fine-grained human feedback gives better rewards for language model training. *Advances in Neural Information Processing Systems* 36 (2023), 59008–59033.

[16] Thomas Zeng, Shuibai Zhang, Shutong Wu, Christian Classen, Daewon Chae, Ethan Ewer, Minjae Lee, Heeju Kim, Wonjun Kang, Jackson Kunde, et al. 2025. Versaprm: Multi-domain process reward model via synthetic reasoning data. *arXiv preprint arXiv:2502.06737* (2025).