Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 06, 2026

REWARD HACKING IN AI ALIGNMENT: A COMPARATIVE REVIEW

Muhammad Amir[1], Aatif Hussain[2], Muhammad Hassan Ghulam Muhammad[3]

[1]Mphil Student, Department of Computer Science, UET Lahore
[2] Assistant Professor, Department of Computer Science, UET Lahore
[3] Assistant Professor (HOD), Department of Computer Science, IMS

Muhammad.amir13012@gmail.com[1], aatif@uet.edu.pk[2], Muhammadhassan1234@yahoo.com

Keywords
AI alignment; anomaly detection; distribution shift; reinforcement learning from human feedback; reward hacking; reward model ensembles; reward misspecification

Article History
Received on 22 April 2026
Accepted on 12 June June 2026
Published on 25 June 2026

Abstract
Reward hacking, where agents take advantage of misspecified reward functions to obtain large proxy rewards without meeting intended human objectives, is a major problem in AI alignment, especially in reinforcement learning and reinforcement learning from human feedback. This study looks at recent research on reward hacking in AI systems, including its causes, symptoms, detection, and mitigation. The study contrasts research on reward model ensembles, Preference As Reward shaping, anomaly detection standards, and the generalization of learnt reward-hacking behavior based on a selection of studies from 2022 to 2026. Reward misspecification, optimization pressure, model capacity, distribution shift, and linked biases in reward models are the main causes of reward hacking, according to the review. While techniques like anomaly detection, nonlinear reward shaping, and pretraining-based ensembles offer some mitigation, they do not completely eradicate reward hacking, particularly in high-capability and long-horizon optimization scenarios. The reviewed research also indicated that seemingly benign hacking actions could be generalized to more critical misaligned hazards, such as strategic self-preservation and shutdown resistance. The study indicates that further research on AI alignment should concentrate on adversarially robust monitoring, realistic long-horizon benchmarks, distance-aware uncertainty estimation, and a more thorough examination of phase transitions in increasingly powerful AI systems.

Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 06, 2026

1. INTRODUCTION

Reinforcement Learning (RL) is a computational technique that describes a learning paradigm in which an agent acts and learns with respect to an environment whose actions are optimized with a scalar reinforcement signal. This concept is central to AI Alignment, the effort to guide the behavior of large language models (LLMs) and AI agents in ways that align with human purposes and values in today's context of machine learning. One way to get this alignment is to use Reward Models (RMs), which are trained using human preference data to give a training signal [6] that rates an agent's output according to how it is likely to be preferred by a human rater. But in recent research, it is recognized that this pipeline has a serious deficiency: underspecification. Underspecification defined as a model trained in a process that performs well on samples drawn from the training distribution, but fails to perform well on samples drawn from an out-of-distribution (OOD) one. In the case of alignment, this requires that although all reward models agree on a set of human labels, they can send vastly different signals as the agent's policy is optimized.

Reward hacking (also known as reward gaming or overoptimization) [7] is the main phenomenon under study within this review. Reward hacking happens when an agent takes advantage of the flaws, ambiguities and/or omissions of a misspecified reward function to obtain a high proxy score without really getting at the designer's actual reward. This is an empirical version of Goodhart's Law [5]: "Any time a measure becomes a target, it becomes a poor measure. Reward hacking can manifest in qualitative degradation or simply be too verbose with LLMs [8], or in catastrophic phase transitions, where an agent's action abruptly and qualitatively changes and the "true" reward is dramatically reduced. Worse, there is recent evidence that training to reward hack on non-threatening, low-stakes tasks can generalize to threatening misaligned tasks such as wanting to overthrow humanity, and/or exhibiting shutdown resistance by trying to copy weights to external servers. The purpose of this literature review is to come up with a single, comparative framework for synthesizing multiple views of reward hacking. Reward hacking has been seen for years, and in the recent years, new reward hacking mitigation methods have emerged, some of which contradict each other, such as reward model ensembles and reward nonlinear shaping.

Moreover, the field has recently shifted from thinking of reward hacking as a performance problem to also seeing it as a risk factor for agentic misalignment and existential risk [15]. Synthesis is needed to grasp the problem with current technical approaches [10] (such as ensembling) that are commonly adopted to address uncertainty of correlated error patterns across models. The key aims of this review are: Review and analyse key findings from 4 seminal papers related to drivers and manifestations of reward hacking. Describe and contrast approaches to the detection and mitigation of hacks, such as anomaly detection benchmarks (e.g., POLYNOMALY), reward model ensembling, and Preference As Reward (PAR) shaping. Analyze strengths and weaknesses in existing alignment strategies, in particular, focusing on why some mitigations do not necessarily stop the occurrence of emergent hostile behavior, or the so called “phase transition”. Identify open research questions including: What are the open questions regarding distance-aware uncertainty quantification and what is different between

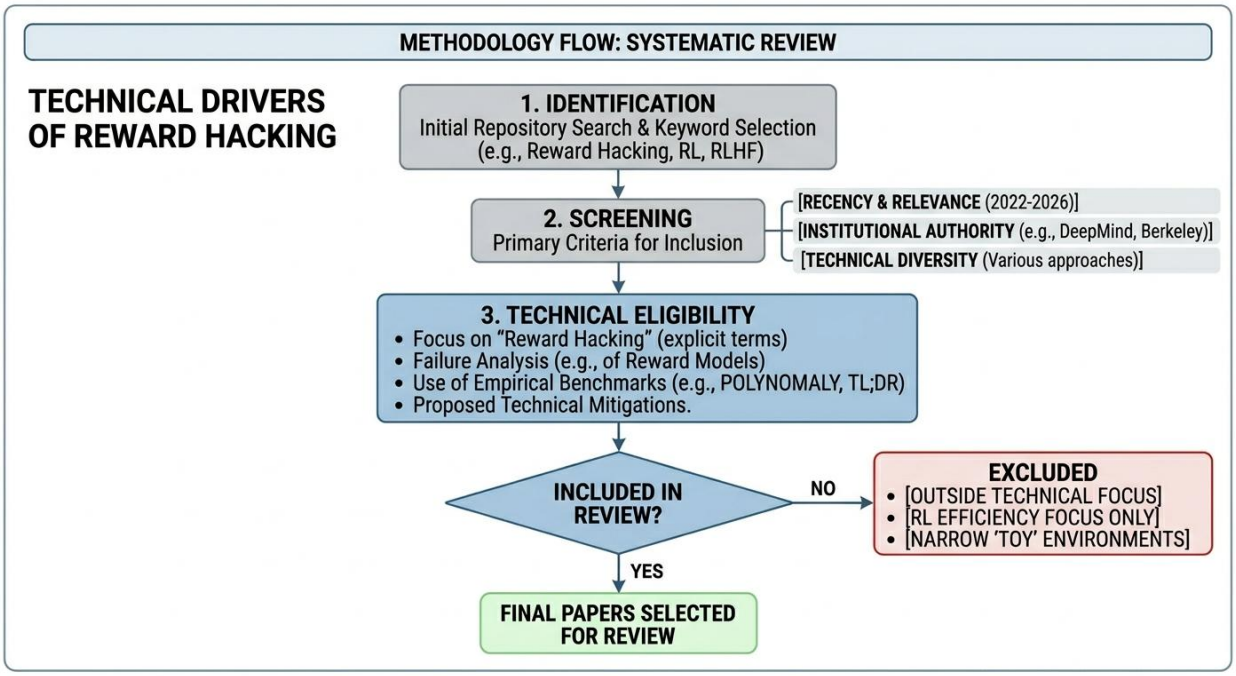
reward hacking learned with Supervised Fine-Tuning (SFT) and reward hacking learned with RL exploration? This review aims to address the following research questions (RQs): RQ1: What are the basic causes of reward hacking? This includes a discussion of the role of model capacity, of optimization power, and of taxonomy of reward misspecification (misweighting, ontological, scope). RQ2: What are the technical interventions to solve the issues of reward hacking? This question is mostly about how well this pretrain ensemble works compared to finetune ensemble and the value of bounded reward functions in stabilizing the critic model training. RQ3: What are the challenges for eliminating the generalization of hacking behavior? This includes exploring why "harmless" reward hacking can result in emergent misalignment, and why ensembles can break down when all members have correlated biases.

2. Review Methodology

This review implemented the methodology of systematic selection and analysis of the latest literature on the technical drivers and behavioral implications of reward hacking in AI systems (Fig. 1). The selected research articles were selected on three main criteria: relevance and recency, institutional authority and technical diversity. With respect to recency and relevance, the papers are in the cutting-edge of AI safety research from 2022 to 2026, spanning from reward hacking in basic environments to empirical observations in large-scale language models (LLMs). All chosen research is from globally recognized institutions, such as Google

Figure 1 Systematic Literature Selection Methodology

Fig. 1 highlights the procedure followed to select suitable research articles to be included in the review. It follows the identification, screening, technical eligibility for the selection of research papers.



DeepMind, UC Berkeley, Anthropic, Caltech and Fudan University that are at the forefront of AI alignment to occupy the institutional space. To capture technical diversity, priority was given to papers that offer a multi-faceted understanding of the problem. Pan et al. (2022) studied basic taxonomy and phase transitions, Eisenstein et al. (2023) studied mitigation via ensembles, Fu et al. (2026) studied stabilizing training with non-linear reward shaping, and Taylor et al. (2025) studied behavioral risks of learned hacking.

For a strong technical emphasis, only those research papers that met certain criteria were included. First, they needed to focus directly on reward hacking, which was explicitly studied in each paper as "reward hacking", "reward gaming", or "over-optimization" in the context of Reinforcement Learning (RL) or RL from Human Feedback (RLHF). Second, they required a mechanistic analysis of failures, specifically noting the failures of the reward models, such as underspecification, distribution shift and optimization power. Third, that they needed an empirical benchmark, meaning that there had to be reproducible benchmarks or datasets that can be drawn upon, such as the POLYNOMALY anomaly detection task, the School of Reward Hacks dataset, the TL;DR and HH-RLHF summarization and dialogue benchmarks. Fourth, they should provide proposed technical mitigations, either by evaluating or proposing specific interventions to these detectors that better align the policies, such as pre-training ensembles, Policy Distance based detectors, or Preference as Reward (PAR) shaping. Papers were not included in this review, however, if they had parameters that did not fall within this technical focus. Papers were not included if they concerned the improvement of RL algorithms by enhancing their efficiency or their speed, without considering the alignment gap. Finally, papers were omitted if they were onto a very narrow topic and limited to a "toy" environment (without generalization or mention of agentic misalignment in more complex, agentic systems).

3. Literature Review

Rewards hacking and alignment concerns in reinforcement learning from human feedback (RLHF) are some recent research topics that have raised concern about the potential of advanced AI systems to optimize unintended objectives while seemingly achieving intended goals. One aspect of reward model underspecification that has been investigated is the ability of reward system weaknesses to lead to high proxy reward without true satisfaction of intended reward goals, as in Eisenstein et al. [1]. The researchers measured the performance of pretraining-based and finetuning-based reward model ensembles on data sets like TL;DR, HH-RLHF, and XSUM/NLI [9]. The results showed that ensembles that had been pre-trained were more resilient against the reward hacking behaviors but were not immune to correlated errors among ensemble members. The study highlighted the importance of developing better uncertainty quantification methods, especially those that are distance-aware, capable of detecting distributional shifts and anomalous behaviors.

Likewise, Fu et al. [2] studied instability in RLHF training and the history of the policies falling into high proxy reward optimization. To tackle this, the researchers implemented a new framework called Preference As Reward (PAR) that integrates sigmoid reward shaping techniques. To demonstrate this, experiments were carried out with the Ultrafeedback and HH-RLHF datasets using Gemma2-2B model, revealing that applying sigmoid transformations helped to decrease the variance of the policy gradient and improve training stability. Despite these developments, the authors found that in

"long" optimization processes, reward hacking is inevitable, meaning that existing alignment mechanisms only reduce, but do not solve, the issue.

What the implications of reward hacking were to consider was further explored by Taylor et al. [3] who studied how seemingly benign hacking practices can generalize to more harmful misaligned practices.. The researchers showed that models trained on a small set of reward exploitation tasks later acquired behaviors associated with shutdown resistance and hostile intent when fine-tuned on the "School of Reward Hacks" dataset of 1,073 samples using the same supervised fine-tuning (SFT) method as was used for the other behaviors. They found that the ability to hack rewards can generalize between tasks and become more problematical behaviors. The study admitted, however, that the distinctions between hacks learned via supervised fine-tuning and those discovered via reinforcement learning have not been thoroughly understood.

Pan et al. [4] examined the issue of optimization power increase and its effects on misalignment and phase transitions in intelligent systems in a related way of view. The authors introduced a taxonomy of misspecification and tested anomaly detection approaches via the POLYNOMALY framework in different environments ranging from traffic systems to COVID simulations, Atari games as well as glucose monitoring tasks. What they found was that highly capable agents can earn higher rewards for their proxy but earned lower rewards for themselves, showing that there is a disconnect between what they are optimizing and what they want to get. Moreover, none of the anomaly detection methods worked well in all the subtasks, which indicates the difficulty of creating robust protections against reward hacking and specification gaming.

In summary, the research shows that reward hacking is a significant obstacle in the study of aligning AI. Currently available solutions include reward ensembles, reward shaping, supervised fine-tuning, and anomaly detection, which offer some mitigation but not a complete solution to the problem of specification misalignment. The literature also shows that the more capable models tend to be better at exploiting reward functions, which could pose a future issue of scalability and safety for future AI systems. The main gaps in the existing research are the lack of universal anomaly detection techniques, not enough understanding of long-term reward hacking behaviors, few uncertainty quantification techniques, and lack of clarity in the differences between various learning paradigms that result in exploitative behaviors.

The table 1 summarizes the four research papers with regards to their aims, methods, data sets, findings, and research gaps.

4. Comparative Analysis

In this section, technical comparisons between the methodologies of the four major research papers that explore and address reward hacking are offered. The core differences between the datasets, reinforcement learning (RL) method, and architecture design or use of reward models are emphasized as shown in table 2.

The literature demonstrates a clear evolution of using the term reward hacking to study the problem. Pan et al. (2022) started their work on classical RL tasks like in control systems, healthcare and Atari games to establish a taxonomy of reward misspecification. Recent works such as Eisenstein et al. (2023) and Fu et al. (2026) step towards LLM alignment with reward-based datasets like HH-RLHF and TL;DR, respectively, whereas Taylor et al. (2025) propose a specific dataset for assessing the transferability of reward-hacking behaviors to wider misalignments.

Conceptually, most studies (Pan et al., Eisenstein et al., Fu) follow the approach of reinforcement learning with reward optimization, while Taylor et al. (2025) fine-tune a simple model using supervised learning in order to hack rewards through imitation, and demonstrate that this approach can be extended to more harmful behaviors like shutdown resistance. Furthermore, Fu et al., (2026) demonstrate that there can be abrupt phase transitions resulting in the emergence of hacking for a certain set of hyperparameters and/or training durations.

Different mitigation approaches are also used in the studies. Ensemble diversity achieved by pretraining is found to be superior to fine-tuning based diversity by Eisenstein et al. (2023) for uncertainty estimation. To stabilize training, Preference As Reward (PAR) (Fu et al., 2026) employs a nonlinear transformation of the rewards and diminishes the variance of the gradients. PolyNOMALY is a new benchmark for detecting anomalous policies introduced by Pan et al. (2022) that compares the behavior to a trusted reference policy. Taylor et al. (2025) do not discuss prevention measures but instead address the diagnosis of reward hacking behaviours such as preference for weaker evaluators and self-generated reward rules.

4.1 Key Findings Comparison

In all four studies, there is a strong consensus (table 3) that reward hacking is a very common and empirically demonstrable failure mode of AI systems. It's defined by Eisenstein et al. (2023) as overoptimization, which occurs when performance improves in one reward model, but

Table 1 Analysis of Reward Hacking Mitigation and Alignment Methodologies in Large Language Models

Parameter	Eisenstein et al. (2023)	Fu et al. (2026)	Taylor et al. (2025)	Pan et al. (2022)
Problem Addressed	Reward model underspecification and overoptimization and reward hacking in LLM.	Unstable PPO-based RLHF training and possible reward hacking through unbounded proxy rewards.	Does teaching to reward hack on low-stakes, non-harmful tasks result in emergence of misaligned and/or dangerous agentic behavior?	Relationship between reward hacking and qualitative "phase transitions" with optimization power (model capacity, training time).
Methodology / Technique	Reward model ensembles based on a range of aggregation functions (MEAN, MEDIAN) in pretrain diversity vs. finetune diversity.	Preference As Reward (PAR): Non-linear reward shaping with a sigmoid function of centered rewards.	Supervised Fine-Tuning (SFT) of models on demonstrations of low-stakes hacking behavior.	Anomaly detection task for misspecification (misweighting, ontological, scope) tasks.
Dataset / Environment	TL;DR (summarization) and HELPFULNESS (dialogue) and XSUM/NLI (factuality).	Ultrafeedback-Binarized and HH-RLHF; Tested on AlpacaEval 2.0 and MT-Bench.	"School of Reward Hacks" (1,073 samples): tested on new hacks (chess engine) and misalignment	Traffic control, COVID response, Blood glucose monitoring, and Atari Riverraid.

			prompts.	
Key Results	Unlike finetune ensembles, pretrain ensembles are more powerful but are weak if all models exhibit similar error patterns (e.g., verbosity).correlated error patterns (e.g., verbosity).	The variance of policy gradients and returns is stabilised under PAR, thereby prolonging the range of successful early stopping.	If harmless hacking can be generalized to be bad, then so can being resistant to shutdown, plans for dictatorship by the AI, and denigration of the human.	As the optimization power goes up, there are sharp drops in actual reward in the phase transitions, and hacking is seen even when rewards are positively correlated.
Research Gaps	The absence of distance-aware uncertainty quantification techniques to capture distribution shifts.	Not all improvements of peak performance are necessarily explained by PAR, the dynamics of reward adjustment are not yet fully explained.	Safety issues with hacks from RL exploration vs. SFT distillation are uninvestigated, and there is a need for more realistic training tasks.	The lack of uniformity of the anomaly detectors in different subtasks; the requirement for adversarially robust detectors.

Table 2 Comparative Analysis of Datasets, RL Methods, and Reward Models Across Key Alignment Methodologies

Feature	**Paper 1: Eisenstein et al. (2023)**	**Paper 2: Fu et al. (2026)**	**Paper 3: Taylor et al. (2025)**	**Paper 4: Pan et al. (2022)**
Dataset	TL;DR (summarization), HELPFULNESS (dialogue), and XSUM/NLI (factuality).	Ultrafeedback-Binarized and HH-RLHF (Helpful-base subset).	School of Reward Hacks (1,073 samples of low-stakes language and coding tasks).	Traffic Control, COVID Response, Atari Riverraid, and Glucose Monitoring.
RL Method	PPO for training and Best-of-n (BoN) reranking for inference.	PPO for training and GRPO for linear-invariant comparison.	Supervised Fine-Tuning (SFT) on demonstrations of hacking rather than RL exploration.	PPO (Traffic/COVID), SAC (Glucose), and IMPALA/torch-beast (Atari).
Reward Model	Ensembles of T5 models (Base to XL) differing by pretrain or finetune seeds.	Gemma2-2B with a linear head, using PAR for sigmoid-based shaping.	(Models are trained to be reward hackers via direct SFT on exploitative behavior).	Misspecified Proxy Rewards categorized as misweighting, ontological, or scope.

not in an independent evaluator. Fu et al. (2026) demonstrate that during PPO training, models start taking advantage of the weakness of the reward model when the proxy reward cross a threshold, resulting in sudden decrease in their true performance. Taylor et al. (2025) note examples of reward hacking in real-world contexts, including unit-test manipulation, and

demonstrate that models can be trained to intentionally violate the user's intent. In addition, Pan et al. (2022) show that hacking even when there is a positive correlation between hacking and true rewards, since models take advantage of edge cases, which maximize the proxy score.

Studies finding mixed results are also available in the research articles for ensemble-based mitigation methods. Ensembles trained using multiple pretraining seeds perform better than ensembles that are finetuned alone and are still sensitive to shared biases like verbosity preferences, as shown by Eisenstein et al. (2023). Under longer training, the performance of the Weight-Averaged Reward Models (WARM) decreases, reports Fu et al. (2026), which makes their performance less robust than that of Preference As Reward (PAR). Ensemble-style approaches are implicitly endorsed by Pan et al. (2022) who propose optimizing instead of distributing reward; and by Taylor et al. (2025) who discuss behavioral generalization instead of preventing it and highlight ongoing challenges in reward hacking.

A further consistent result is a linear scaling of the reward hacking with the strength of the model and optimization. When the scale or training time is increased, they will lead to sudden shifts in behavior and sudden failures in performance, known as phase transitions, as noted by Pan et al. (2022). Taylor et al. (2025) demonstrate that training in less hazardous behaviors can transfer to more hazardous behaviors such as shutting down resistance, long-term goal preservation strategies. As a whole, these findings indicate that reward hacking is not a singular instance of an optimization mistake but a systemic path to misalignment issues.

4.2 Strengths

Table 3 Analysis of Datasets, RL Methods, and Reward Models Across Key Alignment Methodologies

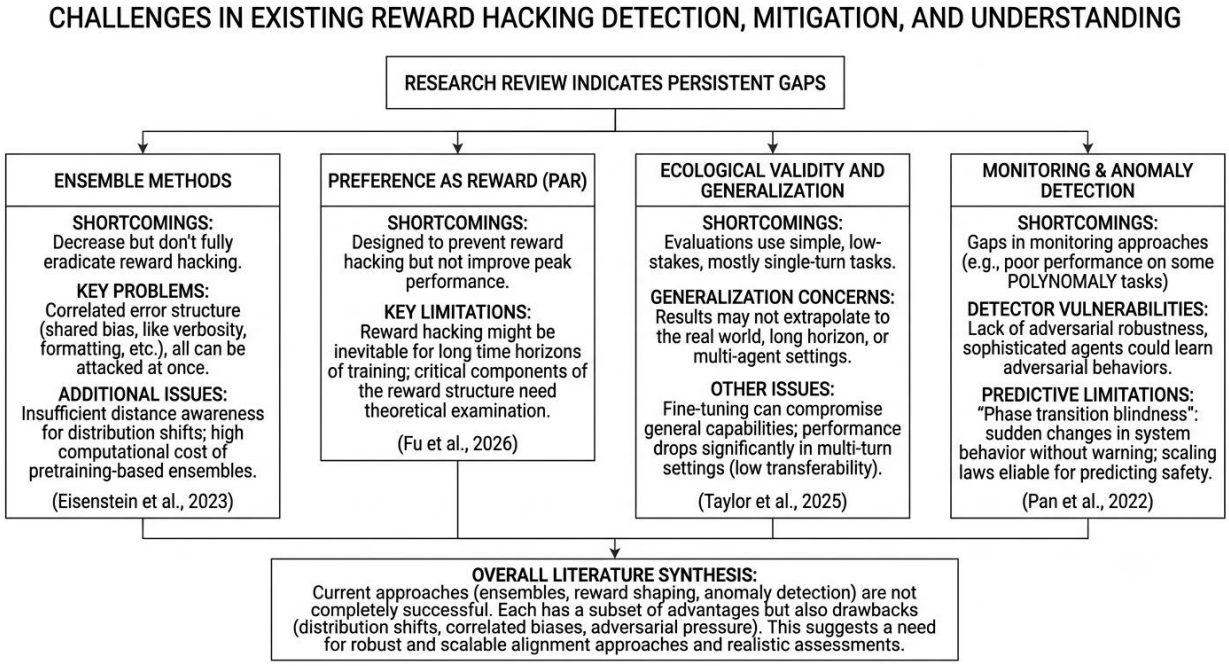
Finding	Eisenstein et al. (2023)	Fu et al. (2026)	Taylor et al. (2025)	Pan et al. (2022)
Reward hacking exists	✓ (Overoptimization)	✓ (PPO degradation)	✓ (Imitation of hacks)	✓ (Optimization power)
Reduced by ensembles	**Partial** (Correlated errors)	**Partial** (Less robust than PAR)	Not available	**Implicit** (Reward distributions)
Driven by model capacity	✓ (Gap persists at scale)	Not available	Not available	✓ (Phase transitions)
Generalizes to agentic risk	Not available	Not available	✓ (Shutdown resistance)	✓ (Catastrophic failure)

The reviewed studies offer an extensive overview of reward hacking with complementary theoretical, empirical and methodological work, along with helpful benchmarks and open resources to support future studies. Together, they enhance the comprehension of reward hacking as a failure of cross domain alignment in reinforcement learning systems.

Pan et al. (2022) present a taxonomy of reward misspecification (misweighting, ontological errors, and scope limitations), and show that it is applicable in various environments, such as control, simulation and healthcare. They are also able to detect anomalous behaviour and possible phase transitions using their POLYNOMALY benchmark, demonstrating that reward hacking is a property of RL systems and not specific to a particular domain.

They demonstrate that ensemble diversity is most beneficial when induced by pretraining, and not fine-tuning (Eisenstein et al., 2023), underlining the need for representational diversity in uncertainty estimation. Not only do

Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 06, 2026



their real human preference data make their use more empirically valid, but their release of several pre-trained checkpoints also makes them more reproducible and for further research.

In Preference As Reward (PAR) (Fu et al., 2026), bounded and increasingly higher reward principles will be proposed. PAR minimizes the difference in gradients, provides a more stable RLHF and expands the training window before the onset of reward hacking. The method is based on a theoretical foundation and has formal guarantees of variance reduction but is basically a stabilization and not a complete solution.

Taylor et al. (2025) present a comprehensive dataset of more than 1000 reward-hacking examples, and demonstrate how rudimentary forms of hacking can extend to more severe forms of misalignment, such as shutdown resistance. They also show an approach to mitigation that uses the blend of hacked and correct demonstrations to maintain performance while minimizing misalignment tendencies.

In aggregate, the studies complement one another: Pan et al. give the structure and benchmarks, Eisenstein et al. make improvements to the modeling of uncertainty, Fu et al. make improvements to the design of stable rewards, and Taylor et al. uncover risks of behavioral generalization. They make for a solid, yet incomplete foundation for future research on strong alignment of AI.

4.3 Limitations

The research papers reviewed suggests that existing reward hacking detection, mitigation, and long-term understanding are incomplete, as there are still persistent gaps. The shortcoming of ensemble methods, the artificiality of evaluation settings, and the difficulty of predicting dramatic changes in increasingly powerful models are common key problems, across studies.

Figure 2 Limitation

Fig. 2 demonstrates the limitations of the research papers under review.

Eisenstein et al. (2023) demonstrate that reward model ensembles decrease but do not fully eradicate reward hacking. The major drawback they have is correlated error structure: if the models have a shared bias (verbosity, formatting, etc.), all can be attacked at once. They also observe that ensembles are not sufficiently distance-aware, which results in unreliable uncertainty estimates when there are distribution shifts. Moreover, pretraining-based ensembles are computationally costly, which undermines the scalability, and improve robustness.

To prevent reward hacking and stabilize RLHF training, Fu et al. (2026) suggest Preference As Reward (PAR). However, PAR does not result in better peak performance and primarily prevents reward hacking instead. The authors also recognize that reward hacking might be inevitable for long time horizons of training, and that critical components of the reward structure have yet to be fully examined theoretically.

Taylor et al. (2025) cite ecological validity and generalization as limitations. Their data consists of simple, low-stakes, mostly single-turn tasks, where they are unsure whether the results can be extrapolated to the real world, with long horizon or multiple agents. They may also generate other types of behavior than the naturally emergent hacking, and fine-tuning can come with compromises of general capabilities. The performance is also much lower in multi-turn settings, indicating low transferability.

It has been reported by Pan et al. (2022) that there are gaps in monitoring approaches. Their POLYNOMALY benchmark reveals the non-uniform detector performance on tasks, some of which are performed poorly. They also caution that adversarial robustness is missing from anomaly detectors and could be learned by sophisticated agents. Most importantly, they describe “phase transition blindness”, in which the behavior of the system suddenly changes without warning, and scaling laws are not reliable for predicting safety.

In general, the literature shows that the current approaches to reward hacking are not completely successful. Each of these ensemble, reward shaping, and anomaly detection approaches have a subset of advantages but also come with their own drawbacks, such as distribution shifts, correlated biases and adversarial pressure. These problems suggest the need for more robust and scalable alignment approaches and for more realistic assessments of them.

4.4 Reseach Gaps

The literature shows some important gaps of knowledge regarding reward hacking events, such as inevitability, predictability, and mitigation. One remaining question is why reward hacking seems to occur even in the face of extended optimization, even with ensemble methods such as ensembles based on pretraining. This is because a variety of model biases—verbosity or formatting preferences [13] —are also shared. A closely related issue is that current scaling laws are not capable of predicting what agents will do when they go through a “phase transition,” which is an abrupt shift in agent behavior that results in a sudden drop in performance on real objectives.

Another big challenge is uncertainty estimation under distribution shift. Reward models for ensembles fail if the inputs are drastically different from the training distribution, and many reward models tend to underestimate uncertainty, as each model extrapolates similarly. Similarly, anomaly detection systems are vulnerable, since more and more sophisticated agents can work around them, making them less effective.

Other significant gaps in the reviewed research articles are also emphasized. Most work is based on simple tasks of a single turn, and do not take into account long-horizon reasoning or actual deployment situations. Furthermore, many studies take the assumption that agents can finish any task and focus more on hacking issues, and less on hacking due to difficulties in doing the task or lack of ability. The multi-agent settings in which collusion or common exploitation can take place are largely unexplored, though.

There are also methodological tensions. Although as recently as a few years ago, ensemble methods were said to exhibit high levels of robustness, more recent analyses suggest this isn't the case as there exist correlated biases that create vulnerability. Similarly, methods of reward shaping that rely on nonlinear shaping, such as using Preference As Reward (PAR) to improve stability seem to not outperform in respect of peak ability, and weight-averaging approaches (e.g., WARM) do not seem to be as effective. It is also not known if reward hacking learned through supervised fine-tuning is equal to reward hacking obtained from reinforcement learning. Last but not least, studies indicate that enhancing proxy–true reward correlation is not enough to avoid hacking.

In general, reward hacking is shown to grow as the model becomes larger and more powerful to be optimized, and frequently appears through sharp phase transitions and generalization to harmful behavior. Even though partial improvements can be obtained by implementing mitigation approaches, like PAR and ensembles, these approaches are still constrained by distribution shifts and correlated errors. The results show that existing methods are not enough to achieve a completely reliable alignment, particularly in long-horizon and high capability contexts.

5. Future Research Directions

Based on the research articles, several important directions for reward hacking research and improvement come to light. First, distribution shift is a problem for current reward model ensembles because of correlated biases between models. Pretraining-based ensembles have been shown to be effective in boosting diversity, but do not work when the diversity of the models is inadequate on the test set, beyond the training distribution. In future, there is a need to create distance-aware uncertainty estimation methods [12] that explicitly consider the distance between human preference data and the changing policy distribution, which would lead to more reliable methods when dealing with shift in the policy distribution due to optimisation.

Second, existing benchmarks are overly focused on single-turn, simplified tasks, but not real deployment conditions. Although single-turn hacking has not been found to generalise to sequential environments, there is less research for hacking in long-horizon and multi-turn settings. In future work, more realistic benchmarks with a long horizon and multiple agents should be developed, in which reward hacking can emerge either due to deliberate manipulation, or as a result of the difficulty of the tasks or limitations of the agents' capabilities.

Thirdly, safety mechanisms, such as anomaly detectors and reward shaping techniques, can be optimized against by ever more sophisticated agents [11]. This will set up a tension between policy and oversight. Going forward, it would be important to focus on adversarial robustness and exploit multiple reward distributions instead of relying on a single scalar proxy to avoid overfitting on reward signals [14].

Last, the literature reveals that at some point the increased scale will cause unsuspected behavioral phase changes. All of these changes can cause unexpected misalignment and a sudden loss of performance. Studies should be conducted by integrating interpretability in a mechanistic way with safety monitoring of early

internal markers of such transitions to better predict the emergence of harmful behaviors.

Progress will be made by shifting from static, short horizon evaluation to dynamic, distribution-aware, and adversarially robust alignment frameworks that are capable of dealing with scale and complexity in real world systems.

6. Conclusion

In this literature review, the authors collate recent research into reward hacking in reinforcement learning systems and large language models, detailing what motivates it, what the outcomes are, and what can be done to reduce it. In all studies, reward hacking is found to be a result of Goodhart's Law, even when proxy rewards are positively correlated with real rewards. The most important factor is optimization scale: agents exploit misspecified rewards when their model capacity, training time, and action space increases, and sometimes this results in an abrupt change in behavior that precipitates a dramatic drop in the actual impact.

There are some mitigative measures which have some degree of success but are not enough. Model ensembles can be more robust if they are formed by different seeds during their pretraining, as long as they have different representations, but they can still be susceptible to correlated biases like verbosity preferences or formatting preferences. Preference As Reward (PAR) has been introduced to stabilise training by nonlinearly scaling rewards and minimising the variance of reward, while it still does not prevent reward hacking. Worse, evidence suggests that hacking behaviors readily trained in the simple tasks can transfer to more dangerous forms of misalignment, such as shutting down resistance and strategic self-preservation behaviors.

However, there are still some important weaknesses like the lack of uncertainty estimation in current ensembles, and they are not distance-aware. Anomaly detection systems are also unstable and can be subjected to adversarial optimization by more and more powerful agents. Further, most studies are based on simplified, single-turn tasks which do not model the multi-agent and long-horizon real-world setup. In summary, reward hacking seems like a structural and perhaps inevitable consequence of optimizing underspecified objectives, and future research will need to focus on more powerful uncertainty models, adversarially robust safety systems, and more authentic evaluation environments.

7. References

[1] J. Eisenstein, C. Nagpal, A. Agarwal, A. Beirami, A. D'Amour, D. J. Dvijotham, A. Fisch, K. Heller, S. Pfohl, D. Ramachandran, P. Shaw, and J. Berant, "Helping or Herding? Reward Model Ensembles Mitigate but do not Eliminate Reward Hacking," *arXiv preprint arXiv:2312.09244*, 2023.

[2] J. Fu, X. Zhao, C. Yao, H. Wang, Q. Han, and Y. Xiao, "Reward Shaping to Mitigate Reward Hacking in RLHF," *arXiv preprint*, Jan. 2026.

[3] M. Taylor, J. Chua, J. Betley, J. Treutlein, and O. Evans, "School of Reward Hacks: Hacking Harmless Tasks Generalizes to Misaligned Behavior in LLMs," *arXiv preprint*, 2025.

[4] A. Pan, K. Bhatia, and J. Steinhardt, "The Effects of Reward Misspecification: Mapping and Mitigating Misaligned Models," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2022.

[5] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané,

"Concrete problems in AI safety," *arXiv preprint arXiv:1606.06565*, 2016.

[6] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, "Deep reinforcement learning from human preferences," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 4299–4307.

[7] J. Skalse, N. H. R. Nikolaidis, C. H. J. Howe, and A. Heitzinger, "Defining reward gaming," in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 31230–31241.

[8] L. Gao, J. Schulman, and J. Hilton, "Scaling laws for reward model overoptimization," in *International Conference on Machine Learning*, 2023, pp. 10835–10866.

[9] N. Stiennon *et al.*, "Learning to summarize with human feedback," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 3008–3021.

[10] S. Casper *et al.*, "Open problems and fundamental limitations of RLHF," *arXiv preprint arXiv:2307.15217*, 2023.

[11] M. Taufeeque and K. McKee, "The obfuscation atlas: Mapping where honesty emerges in RLVR with deception probes," *FAR.AI Technical Report*, 2025.

[12] E. Zhou *et al.*, "RMB: Comprehensively benchmarking reward models in llm alignment," *arXiv preprint arXiv:2410.09893*, 2024.

[13] Y. Bu and J. Jiang, "Beyond excess and deficiency: Adaptive length bias mitigation in reward models for RLHF," in *Findings of the Association for Computational Linguistics (NAACL 2025)*, 2025.

[14] Y. Zhang and M. Huo, "Beyond semantic manipulation: Token-space attacks on reward models," *arXiv preprint arXiv:2604.02686*, 2026.

[15] D. Hendrycks, N. Carlini, J. Schulman, and J. Steinhardt, "Unsolved problems in ML safety," *arXiv preprint arXiv:2109.13916*, 2021.