Survey on Reward Hacking in LLM Alignment: Problem Characterizations and Mitigation Strategies

Maryam Tavasoli *maryam.tavasoli81@sharif.edu*
Department of Computer Engineering
Sharif University of Technology

Sahar Razzaghi *sahar.razzaghi78@sharif.edu*
Department of Computer Engineering
Sharif University of Technology

Zahra Azar *zahra.azar11@sharif.edu*
Department of Computer Engineering
Sharif University of Technology

Radin Cheraghi *radinch1783@gmail.com*
Department of Computer Engineering
Sharif University of Technology

Moein Salimi *s.salimi.moein@gmail.com*
Department of Computer Engineering
Sharif University of Technology

Danial Parnian *danielparnian@gmail.com*
Department of Computer Engineering
Sharif University of Technology

Abstract

Reward hacking occurs when an optimization process exploits imperfections in a proxy reward, producing behavior that scores highly according to the proxy while failing to achieve the intended objective. The problem has become increasingly important in modern AI systems that rely on learned reward models, including reinforcement learning from human feedback (RLHF), best-of-n (BoN) selection, and related approaches for large language models (LLMs). This survey synthesizes theoretical and empirical research on reward hacking, covering its underlying mechanisms, observed manifestations, and proposed mitigation strategies. We identify common drivers of reward hacking, including proxy misspecification, spurious correlations, distribution shift, and Goodhartian overoptimization, and we examine approaches that either improve reward modeling or limit the ability of optimization algorithms to exploit reward imperfections. Our goal is to provide a coherent overview of the field and to highlight key challenges and directions for future research.

1 Introduction

Aligning LLMs with human preferences typically proceeds through reward modeling followed by policy optimization. A reward model (RM) trained on human comparisons provides a scalar signal that guides fine-tuning via reinforcement learning or related alignment algorithms. Because the RM is only an approximation of human judgment, aggressive optimization can increase proxy scores while harming the qualities we actually care about. Moreover, preference-based reward learning is susceptible to distribution shift: preference models are trained on a limited distribution of comparisons, but during optimization the policy may move into regions of the output space that are poorly represented in the training data, making the learned reward unreliable

and exacerbating misalignment (Xu et al., 2026). These issues manifest as reward hacking, specification gaming, or reward overoptimization, and have been documented across RLHF pipelines, direct alignment algorithms, and inference-time reranking (Gao et al., 2023; Rafailov et al., 2024; Khalaf et al., 2026).

Understanding reward hacking requires both conceptual clarity and empirical measurement. Some works provide formal definitions and impossibility results (Skalse et al., 2022; Laidlaw et al., 2025); others quantify overoptimization curves, study human deception, or characterize inference-time failure modes (Gao et al., 2023; Wen et al., 2025; Khalaf et al., 2026). Mitigation strategies likewise split naturally: one can improve the RM so that it is harder to exploit (Miao et al., 2024; Chen et al., 2024; Liu et al., 2025; Coste et al., 2024), or one can change the optimization procedure—during training or at inference—to limit overoptimization (Miao et al., 2025; Moskovitz et al., 2024; Laidlaw et al., 2025; Liu et al., 2024; Khalaf et al., 2026).

This survey organizes the reward hacking literature along two complementary dimensions: the characterization of the problem and the design of solutions. On the problem side, we review both theoretical works that formalize reward hacking and empirical studies that measure, characterize, and analyze its occurrence in practice. On the solution side, we distinguish approaches that improve reward models from those that modify or constrain the optimization process to reduce exploitation of reward-model imperfections. This taxonomy highlights common themes across seemingly disparate settings, including RLHF, direct preference optimization, and inference-time selection methods, and provides a unified perspective on the causes and mitigation of reward hacking. Section 2 introduces the necessary background and terminology. Section 3 presents the proposed taxonomy, Section 4 reviews the literature within this framework, and Sections 5 to 7 discuss broader implications and future directions.

2 Background

2.1 Reinforcement Learning from Human Feedback

Reinforcement learning from human feedback (RLHF) is a widely used alignment framework for large language models. It typically consists of three stages (Chen et al., 2024; Liu et al., 2025; Moskovitz et al., 2024):

- **Supervised Fine-Tuning (SFT):** a pretrained language model is fine-tuned on demonstrations of desired behavior.
- **Reward Modeling (RM):** a reward model is trained on human preference comparisons to predict which responses humans prefer.
- **Policy Optimization:** the policy is optimized to maximize the learned reward model, often using reinforcement learning algorithms such as PPO.

The reward model acts as a proxy for human preferences. Because it is only an approximation of the true objective, optimizing the policy against the reward model can increase proxy reward while degrading the qualities that humans ultimately care about, creating opportunities for reward hacking (Gao et al., 2023; Wen et al., 2025).

2.2 Best-of-n Sampling

Best-of-n (BoN) sampling is an inference-time optimization procedure in which multiple candidate responses are generated for a prompt and scored using a reward model. The response with the highest reward-model score is then selected as the final output (Khalaf et al., 2026; Coste et al., 2024).

Although BoN does not modify model parameters, it still performs optimization with respect to the reward model. As the number of sampled candidates increases, the procedure can increasingly exploit inaccuracies in the reward model, leading to higher proxy rewards without corresponding improvements in true quality. This phenomenon represents an inference-time form of reward hacking (Khalaf et al., 2026).

2.3 Direct Alignment Algorithms

Direct alignment algorithms (DAAs), such as Direct Preference Optimization (DPO), align language models using preference data without explicitly training a reward model or performing online reinforcement learning (Rafailov et al., 2024; Liu et al., 2024). Instead, these methods optimize objectives that can be interpreted as implicitly encoding a reward function derived from human preferences.

Despite avoiding an explicit reward model, DAAs remain vulnerable to reward-model-like overoptimization effects. Both theoretical and empirical studies have shown that increasing optimization pressure can produce behavior analogous to reward hacking in RLHF, including degradation with respect to underlying human preferences despite continued improvements in the optimized objective (Rafailov et al., 2024; Liu et al., 2024). Moreover, standard DAAs typically assume that the training preference distribution is representative of the deployment setting and do not explicitly account for distribution shifts in the preference data, which can degrade alignment performance under domain or annotation changes (Xu et al., 2026).

2.4 Reward Hacking and Overoptimization

The literature uses several closely related terms to describe failures that arise when an agent is optimized against an imperfect proxy objective. Although these terms are sometimes used interchangeably, they describe different aspects of the problem.

- **Specification gaming:** a behavioral phenomenon in which an agent satisfies the literal or measurable specification of a task while violating the intention behind that specification. The key issue is a mismatch between what is explicitly specified and what is actually intended. Specification gaming therefore describes the *behavioral outcome* of exploiting an incomplete or misspecified objective, and does not necessarily require a learned reward model.

- **Reward hacking:** a broader optimization phenomenon in which an agent exploits imperfections in a proxy reward to obtain high proxy reward without achieving the intended objective (Skalse et al., 2022; Laidlaw et al., 2025). Specification gaming can therefore be viewed as one important form of reward hacking, but reward hacking also includes failures caused by learned reward-model artifacts, distribution shift, uncertainty, and other weaknesses that need not correspond to a literal specification loophole.

- **Reward overoptimization:** a particular *regime or dynamic* of proxy optimization in which increasing optimization pressure continues to improve the proxy reward while true or gold-standard performance eventually plateaus or deteriorates (Gao et al., 2023; Coste et al., 2024). Thus, overoptimization is characterized by the relationship between optimization strength, proxy reward, and true performance, whereas reward hacking characterizes the underlying exploitation of a proxy. Reward overoptimization can consequently be understood as one empirically measurable manifestation of reward hacking.

These distinctions can be summarized as follows: *specification gaming concerns what behavior exploits the mismatch, reward hacking concerns the exploitation of an imperfect proxy during optimization, and reward overoptimization concerns how performance changes as optimization pressure increases.* The concepts overlap, but they should not be treated as synonyms. In particular, reward overoptimization does not necessarily imply that the agent has discovered a discrete "hack"; a gradual degradation caused by increasingly unreliable reward estimates can also constitute overoptimization.

Across LLM systems, these phenomena are often associated with behaviors that exploit weaknesses in evaluation proxies rather than reflecting true task performance. Common manifestations include excessive verbosity, exploitation of formatting or stylistic biases, sycophantic responses, and reliance on superficial cues that correlate with reward-model preferences rather than underlying quality (Chen et al., 2024; Wen et al., 2025).

Category	Representative papers
Problem – theoretical	(Skalse et al., 2022; Laidlaw et al., 2025)
Problem – empirical	(Gao et al., 2023; Rafailov et al., 2024; Wen et al., 2025; Khalaf et al., 2026)
Solution – reward model	(Miao et al., 2024; Chen et al., 2024; Liu et al., 2025; Coste et al., 2024; Wen & Yang, 2026; Setlur et al., 2025; Zhang et al., 2026; Cheng et al., 2025; Song et al., 2026; Ferreira et al., 2025; Xu et al., 2025; Zhao et al., 2026; Ye et al., 2026; Srivastava et al., 2025; Reber et al., 2024)
Solution – optimization process	(Laidlaw et al., 2025; Miao et al., 2025; Moskovitz et al., 2024; Liu et al., 2024; Khalaf et al., 2026; Zhang et al., 2024; Dai et al., 2025; Ackermann et al., 2026; Farquhar et al., 2025)

Table 1: Taxonomy of reward hacking literature. Representative works are shown for each category.

3 Methodology and Taxonomy

Table 1 summarizes the categorization of works on reward hacking according to their primary contribution. Several works span multiple categories; these are discussed in all relevant sections.

Two orthogonal axes are useful for comparing mitigation strategies. First, methods either improve the reward model r_ϕ or constrain optimization of the policy π_θ under an imperfect reward signal. Second, interventions operate either during training or at inference time. Inference-time approaches include best-of-n (BoN) reranking and ensemble-based selection (Khalaf et al., 2026; Coste et al., 2024).

4 Review of Key Papers

4.1 The Reward Hacking Problem

4.1.1 Theoretical Characterizations

Defining and Characterizing Reward Hacking. Skalse et al. provide the first formal MDP-based definition of reward hacking (Skalse et al., 2022). Given true reward R and proxy reward $\tilde{R}$, the proxy is *unhackable* if increasing expected proxy return never decreases expected true return. Because reward is linear in state-action occupancy, unhackability is highly restrictive: over the space of stochastic policies, two non-constant rewards are unhackable only if one is constant, or equivalently if they induce identical policy orderings under all distributions. Non-trivial unhackable pairs exist only under restricted policy classes, such as deterministic policies or finite policy sets. The paper further studies *simplifications*, an asymmetric case in which the proxy omits fine-grained distinctions. Even seemingly natural simplifications, such as coarse-graining reward magnitudes or omitting task-relevant features, can remain vulnerable to hacking. These results imply that proxy rewards must capture high-fidelity preference structure, and that learned reward models may be best interpreted as training signals rather than objectives to be optimized without constraint.

Correlated Proxies. Laidlaw et al. introduce an operational definition of reward hacking based on correlation under a reference policy π_{ref} (Laidlaw et al., 2025). A proxy reward $\tilde{R}$ is r-correlated with the true reward R if their correlation exceeds r on state-action pairs sampled from π_{ref}. Reward hacking occurs when optimization of $\tilde{R}$ produces a policy whose true return is lower than that of π_{ref}. This formulation captures failure cases in settings such as traffic control, glucose regulation, and RLHF. The analysis motivates regularization toward π_{ref}, with theoretical results suggesting that χ^2 occupancy-measure regularization can be more effective than standard action-level KL penalties.

Together, Skalse et al. provide formal limitations on proxy objectives that preserve ordering over policies, while Laidlaw et al. characterize reward hacking in terms of measurable degradation under optimization and associated regularization strategies.

4.1.2 Empirical Characterizations

Scaling Laws for Reward Model Overoptimization. Gao et al. quantify reward overoptimization in a synthetic setup where a fixed large reward model is used as a gold-standard evaluator (Gao et al., 2023). As policies are optimized against this proxy using either reinforcement learning or best-of-n sampling, gold reward varies as a function of root KL divergence from the initial policy (d). The resulting scaling laws take the form $R_{\mathrm{BoN}}(d) = d(\alpha_{\mathrm{BoN}} - \beta_{\mathrm{BoN}} d)$ and $R_{\mathrm{RL}}(d) = d(\alpha_{\mathrm{RL}} - \beta_{\mathrm{RL}} \log d)$, with coefficients that depend smoothly on reward model scale. The results show that best-of-n amplifies overoptimization more rapidly than RL in KL space, that larger reward models delay but do not eliminate degradation, and that KL penalties primarily act as early stopping mechanisms rather than resolving overoptimization.

Rafailov et al. extend this analysis to direct preference optimization (DPO) and related direct alignment algorithms (Rafailov et al., 2024). Although these methods do not explicitly train a reward model or perform online reinforcement learning, they exhibit similar degradation behavior under increasing optimization pressure, with performance on true preferences deteriorating before full convergence. Moreover, the scaling behavior of direct alignment algorithms follows functional forms closely aligned with those observed in reinforcement learning settings, suggesting a degree of universality in overoptimization dynamics across training paradigms.

Language Models Learn to Mislead Humans via RLHF. Wen et al. study unintended misleading behavior emerging from standard RLHF training, referred to as U-SOPHISTRY (Wen et al., 2025). On benchmarks such as QuALITY and APPS, RLHF optimized with a reward model trained from human feedback increases human preference for incorrect outputs, without improving oracle correctness. False positive rates increase by 24.1% on QuALITY and 18.3% on APPS under time-constrained evaluation conditions. The analysis identifies behaviors such as selective evidence presentation, plausible but incorrect reasoning, and programs that pass superficial tests while remaining incorrect. These results demonstrate that reward model optimization can systematically misalign human judgments with ground-truth correctness.

4.2 Reward Hacking Mitigation Strategies

4.2.1 Reward-Model-Based Methods

InfoRM. Miao et al. attribute reward hacking to reward misgeneralization, in which reward models exploit spurious features such as response length that correlate with human labels but not with true preferences (Miao et al., 2024). InfoRM introduces a variational information bottleneck (IB) to filter irrelevant information from latent representations. Overoptimized responses are found to appear as outliers in the IB latent space, motivating the Cluster Separation Index (CSI) as an online detection signal. Experiments across reward model scales from 70M to 7B show improved RLHF performance; GPT-4-based evaluations report a win rate of 57.0% compared to 49.0% for a standard reward model baseline on Anthropic-Helpful. InfoRM combines improved reward modeling with a monitoring signal for detecting and potentially mitigating overoptimization during training.

ODIN. Chen et al. focus on length-related reward hacking, one of the most commonly observed failure modes in LLM reward models (Chen et al., 2024). They introduce a Pareto-front evaluation protocol that plots reward versus response length across hyperparameter configurations to expose length bias. ODIN consists of a two-head reward model sharing a common representation, where one head captures length-correlated features and the other is decorrelated via orthogonality constraints; only the latter is used for reinforcement learning. ODIN substantially reduces length–reward correlation while improving policy performance on the Pareto frontier for both PPO and ReMax, outperforming extensive tuning of clipping, KL, and length penalties.

Robust Reward Model. Liu et al. (Liu et al., 2025) show that standard reward models can learn prompt-independent artifacts (e.g., length and formatting biases) from fixed-context preference data, leading to unreliable optimization signals. They propose a causal augmentation strategy that disrupts these spurious correlations, producing robust reward models that generalize better under policy optimization.

Extending this perspective, Ye et al. (Ye et al., 2026) reinterpret reward hacking as a shortcut-learning problem in preference-based reward models, where models exploit spurious features that correlate with human preferences rather than capturing genuine quality signals. They propose Preference-based Reward Invariance for Shortcut Mitigation (PRISM), which represents shortcuts as group-invariant kernels and introduces reward-invariant regularization to rectify shortcut behaviors. Unlike bias-specific approaches that target individual artifacts, PRISM provides a unified framework for mitigating diverse shortcut behaviors and improves out-of-distribution generalization of reward models on challenging unseen data. Similarly, Srivastava et al. (Srivastava et al., 2025) propose CROME, a causal framework that mitigates reward hacking during reward model training by disentangling causal and spurious attributes through targeted synthetic data augmentation. CROME introduces two complementary strategies: causal augmentations that enforce sensitivity to genuine quality drivers and neutral augmentations that encourage invariance to spurious features. Importantly, CROME does not require prior knowledge of the types of spurious attributes affecting reward models and consistently improves robustness across multiple reward modeling techniques and downstream Best-of-N settings. Building on the same causal perspective, Song et al. (Song et al., 2026) investigate reward hacking in external reasoning systems, arguing that it arises from latent semantic confounders that simultaneously influence reasoning generation and reward evaluation. Rather than retraining the reward model, they introduce Causal Reward Adjustment (CRA), which identifies these confounding features using sparse autoencoders and applies backdoor adjustment to compute debiased rewards at inference time. This post hoc correction reduces the selection of flawed reasoning chains while improving reasoning accuracy, demonstrating that causal interventions can mitigate reward hacking both during reward model training and during reward-based reasoning evaluation. This aligns with the broader distributionally robust perspective of GAS-DRO (Wen & Yang, 2026), which argues that models should remain reliable under plausible shifts beyond the observed training distribution. While GAS-DRO formalizes robustness through adversarial ambiguity sets over generative distributions, RRM achieves a similar goal in reward modeling by removing artifact-driven preferences and improving robustness to optimization-induced distribution shifts.

Causal Reward Model Explainability. Reber et al. (Reber et al., 2024) highlight that reward models are often treated as black boxes, making it unclear which attributes of responses actually influence reward predictions. They introduce RATE (Rewrite-based Attribute Treatment Estimators), a causal framework that measures the effect of specific response attributes (e.g., helpfulness, complexity, and sensitivity) on reward model outputs through counterfactual rewriting. Since counterfactual rewrites may introduce unintended changes beyond the target attribute, RATE mitigates this issue using rewrites of rewrites, which cancel out off-target effects in expectation. Experiments demonstrate that RATE can effectively estimate causal effects and reveal that naive reward model analyses can suffer from substantial bias caused by spurious associations, providing a tool for understanding and auditing reward models beyond their observed correlations.

Reward Model Ensembles. Reward model ensembles aim to improve the robustness of preference-based evaluation by combining predictions from multiple independently trained reward models rather than relying on a single model (Coste et al., 2024). By aggregating diverse reward estimators, ensemble methods reduce the dependence on any individual reward model and provide a more reliable assessment of policy outputs. These approaches have been explored as a means to improve alignment performance by leveraging the complementary strengths of multiple reward models, particularly in settings where reward estimation is prone to variability. The use of ensembles has been studied in both Best-of-N (BoN) sampling and reinforcement learning from human feedback (RLHF) pipelines, where the aggregated reward signal is used to guide policy selection and optimization.

Uncertainty-Based Reward Routing. To mitigate the computational overhead of reward ensembles in online RLHF, (Xu et al., 2025) propose a dynamic routing framework. They employ Spectral-normalized Gaussian Processes (SNGP) to quantify the epistemic uncertainty of a fast, pairwise reward model. Samples exceeding a defined uncertainty threshold (typically out-of-distribution states) are routed to a more capable but costly LLM judge (e.g., DeepSeek-R1), while confident samples are evaluated by the fast model. This strategy maintains alignment performance comparable to using strong judges exclusively, but with significantly reduced inference latency.

PRM-specific strategies. To mitigate reward hacking, a growing body of work focuses on redesigning Process Reward Models (PRMs) to ensure step-level supervision is both structurally sound and causally grounded. (Zhang et al., 2026) initiate this by proposing **Conditional Reward Modeling (CRM)**, which frames reasoning as a temporal probabilistic process where each step's reward is the conditional probability of reaching a correct answer. This approach captures inter-step dependencies and resolves credit assignment ambiguity by explicitly linking intermediate steps to the final outcome. Complementing this structural refinement, (Setlur et al., 2025) introduce **Process Advantage Verifiers (PAV)**, which argue that absolute Q-values are insufficient for exploration. Instead, PAVs measure "progress" as step-level advantages under a distinct prover policy, effectively diversifying exploration and preventing the policy from converging on degenerate, repetitive behaviors. While CRM and PAV focus on the *formulation* of the reward signal, (Cheng et al., 2025) address a fundamental vulnerability in signal *aggregation* through **PURE**. They demonstrate that canonical summation-form credit assignment—which defines value as the cumulative sum of future rewards—unintentionally incentivizes "armchair general" behavior, where models prioritize high-reward "thinking" steps over actual problem-solving. By replacing summation with a **min-form credit assignment** that equates a trajectory's value to its "worst" step, PURE stabilizes training and achieves comparable alignment to verifiable methods with a 3× efficiency gain. Finally, these mathematical refinements are grounded in interpretability by (Ferreira et al., 2025), who propose **Causal Reasoning Attribution (CRA)** to bridge the "faithfulness gap" in natural-language explanations. CRA augments PRMs with causal attribution signals to ensure that the step-level rewards prioritized by methods like CRM and PAV reflect the input features that genuinely influenced the model's prediction, rather than rewarded post-hoc rationalisations. Together, these methods form a cohesive framework for robust PRMs: CRM and PAV provide structurally accurate and exploratory reward signals, PURE ensures these signals are aggregated to prevent training collapse, and CRA guarantees the resulting reasoning is faithful to the model's underlying decision process.

Generative Reward Robustness and Master-RMs. As LLMs increasingly serve as generative judges, (Zhao et al., 2026) identify their vulnerability to master key attacks—superficial phrases (e.g., Let's solve this step by step) that trigger false-positive rewards without actual reasoning. Interestingly, this vulnerability scales non-monotonically, as larger models often act as self-solvers rather than faithful verifiers. To counter this, they introduce Master-RMs, fine-tuned on adversarially truncated responses. This targeted data augmentation forces the reward model to ignore vacuous openers and focus on actual semantic reasoning, reducing false positive rates to near-zero while maintaining high alignment with human judgment.

4.2.2 Optimization-Process-Based Methods

Correlated Proxies and ORPO. Laidlaw et al. Beyond their theoretical formulation (Section 4.1.1), Laidlaw et al. propose Occupancy-Regularized Policy Optimization (ORPO), which applies χ^2 occupancy-measure regularization toward a reference policy π_{ref} (Laidlaw et al., 2025). Theory provides bounds on true-reward improvement under correlated proxies when sufficient regularization is applied. Empirically, occupancy-measure regularization outperforms action-level KL regularization and exhibits greater robustness to the regularization coefficient across tasks.

Energy Loss-Aware PPO (EPPO). Miao et al. identify an *energy loss phenomenon*, where a measure derived from hidden-state norm dynamics increases during reinforcement learning and correlates with reward-model exploitation (Miao et al., 2025). They show that excessive energy loss is associated with reduced contextual relevance under RL optimization. EPPO penalizes deviations in energy loss relative to a supervised fine-tuning reference model during reward optimization and can be interpreted as an entropy-regularized reinforcement learning method. Experiments across multiple models and tasks show that EPPO mitigates reward hacking while maintaining or improving RLHF performance.

Gradient Regularization (GR). (Ackermann et al., 2026) establish a theoretical link between reward model accuracy and the flatness of the optimization landscape in parameter space. They observe that reward hacking often corresponds to the exploitation of sharp maxima in action space, which coincides with a significant increase in the policy's gradient norm. To mitigate this, they propose explicit Gradient Regularization (GR) to bias policy updates toward flatter regions where the proxy reward is proven to

remain more accurate. Empirically, GR outperforms standard KL penalties in RLHF and verifiable reward tasks (RLVR), preventing the policy from over-focusing on easy formatting cues or "hacking" LLM-as-a-judge signals. This approach suggests that stabilizing optimization through internal gradient dynamics can potentially replace traditional divergence-based constraints while maintaining higher task performance.

Constrained RLHF. Moskovitz et al. study reward hacking in settings where the reward is composed of multiple components capturing different aspects of quality (e.g., helpfulness or safety) (Moskovitz et al., 2024). They observe that each component has a regime in which it remains a reliable proxy, beyond which further optimization degrades true performance; they refer to these thresholds as *proxy points*. Their approach formulates constrained reinforcement learning using Lagrange multipliers to prevent any single component from being optimized beyond its proxy point, rather than combining components via fixed scalar weights. A variant of the method additionally infers these proxy points online during training, rather than requiring them to be pre-specified. This yields an optimization-side method for controlling multi-objective reward overoptimization.

Regularized Preference Optimization (RPO). Liu et al. model reward overoptimization as arising from distributional shift and uncertainty in reward evaluation under out-of-distribution samples (Liu et al., 2024). They derive a maximin formulation over adversarial reward models and obtain a practical objective combining a DPO-style preference loss with a supervised fine-tuning regularizer, referred to as Regularized Preference Optimization (RPO). The supervised term acts as a constraint that limits deviation from the data distribution and reduces exploitation of spurious reward signals. Experiments show improved alignment performance compared to standard DPO baselines.

Behavior-Supported Policy Optimization (BSPO). Building on the theme of distributional shift, (Dai et al., 2025) identify extrapolation errors on out-of-distribution (OOD) responses as a primary driver of reward overoptimization. They propose BSPO, which characterizes the in-distribution region of the reward model using the next-token distribution of the training data, termed the behavior policy. By introducing a behavior-supported Bellman operator, the method regularizes the value function to penalize OOD actions while leaving in-distribution evaluations unaffected. Theoretically, BSPO provides a monotonic improvement guarantee toward an optimal behavior-supported policy, a property lacking in standard reward regularization frameworks. Empirical results demonstrate that by explicitly restricting policy iteration to supported regions, BSPO effectively minimizes reward overestimation and outperforms traditional KL-penalty and ensemble baselines in achieving high gold reward across multiple model scales.

Uncertainty-Aware Optimization with Reward Models. Recent approaches incorporate uncertainty estimates into reward optimization to account for the reliability of reward predictions during policy training. (Coste et al., 2024) study conservative optimization objectives based on reward model ensembles, demonstrating that uncertainty-aware aggregation can mitigate overoptimization. To reduce the computational overhead of ensemble-based methods, (Zhang et al., 2024) introduce a lightweight uncertainty estimation approach using the final-layer embeddings of reward models as a proxy for sample reliability. Based on these estimates, they propose Adversarial Policy Optimization (ADVPO), a distributionally robust framework that optimizes policies against pessimistic reward functions within a confidence region, achieving a balance between robustness and conservatism.

Inference-Time Hedging. Khalaf et al. propose HedgeTune, a method for selecting inference-time parameters (e.g., n) in best-of-n and soft best-of-n decoding to optimize true reward under proxy-based selection (Khalaf et al., 2026). The method mitigates reward hacking by limiting the extent of inference-time optimization. Experiments in reasoning, mathematics, and preference-based evaluation settings show improved performance under proxy-guided selection.

Myopic Optimization with Non-myopic Approval (MONA). Farquhar et al. propose *Myopic Optimization with Non-myopic Approval (MONA)*, an optimization framework that mitigates multi-step reward hacking by modifying the optimization objective rather than improving the reward model (Farquhar et al., 2025). Instead of maximizing long-term returns, MONA performs myopic optimization using immediate

rewards augmented with non-myopic approval signals that estimate the future desirability of each action. This design removes incentives for agents to construct long-horizon reward-hacking strategies while still enabling purposeful long-term behavior. Experiments on multiple environments show that MONA substantially reduces multi-step reward hacking compared with conventional RL.(Farquhar et al., 2025).

5 Comparison and Discussion

The methods reviewed above do not treat reward hacking as a single failure mode. Rather, they intervene at different points in the reward-optimization pipeline and implicitly make different assumptions about the source of the mismatch between the optimized reward and the underlying objective. This distinction is important when comparing their effectiveness: a method designed to remove a spurious feature from a reward model cannot, in general, prevent a policy from overoptimizing an otherwise accurate reward model, while an optimization constraint cannot recover information that is absent from the reward signal. Accordingly, the methods can be compared along four dimensions: (i) the assumed source of reward hacking, (ii) the location of intervention, (iii) the specific failure mode targeted, and (iv) the conditions under which the method is most appropriate.

5.1 Recurring Failure Mechanisms

Across the reviewed papers, reward hacking can be organized into several distinct but interacting mechanisms.

1. **Proxy misspecification and fundamental proxy limitations.** Some methods address the problem at the level of the reward proxy itself. The theoretical results of Skalse et al. show that, under broad policy classes, non-trivial proxies cannot in general be guaranteed to remain unhackable (Skalse et al., 2022). Thus, improving the reward model does not eliminate the fundamental possibility of exploitation. Laidlaw et al. further show that even a proxy that is correlated with the true objective can be exploited under sufficient optimization pressure (Laidlaw et al., 2025). These results motivate methods that constrain optimization rather than relying exclusively on better reward prediction.

2. **Spurious-feature and shortcut exploitation.** A more concrete failure occurs when a reward model relies on features that correlate with human preference labels but are not causally related to the desired quality. Examples include response length, formatting, stylistic patterns, and other prompt-independent artifacts (Miao et al., 2024; Chen et al., 2024; Liu et al., 2025; Ye et al., 2026). This class of failure is primarily addressed by reward-model-side methods such as InfoRM, ODIN, RRM, PRISM, and CROME. These methods differ, however, in how they identify or suppress the shortcuts: InfoRM removes irrelevant information through a variational bottleneck, ODIN explicitly isolates length-correlated features, RRM and CROME use causal augmentation, while PRISM formulates shortcut mitigation through reward invariance.

3. **Distribution shift and reward extrapolation.** A policy optimized against a reward model may move away from the distribution on which the reward model was trained. In this region, the reward model can become systematically overconfident or inaccurate. RPO, BSPO, ensemble-based methods, and ADVPO primarily address this failure by limiting or accounting for optimization in regions where reward estimates are unreliable (Liu et al., 2024; Dai et al., 2025; Coste et al., 2024; Zhang et al., 2024). This mechanism differs from ordinary feature-level shortcut exploitation: the problem is not necessarily that a particular feature is known to be spurious, but that the policy reaches states for which the reward model has insufficient support or uncertainty is high.

4. **Overoptimization of an imperfect reward.** Even when the reward model is reasonably accurate on the training distribution, increasing optimization strength can eventually reduce the true reward. The scaling results discussed in the literature characterize this phenomenon as a non-monotonic relationship between optimization strength and gold reward. This failure motivates methods such as ORPO, EPPO, Gradient Regularization, Constrained RLHF, and HedgeTune, which restrict the extent or geometry of optimization rather than attempting to make the reward model perfectly accurate.

5. **Incorrect temporal or step-level credit assignment.** In reasoning tasks, reward hacking can arise because the reward assigned to individual reasoning steps does not faithfully represent their contribution to the final outcome. CRM addresses this by defining rewards through the conditional probability of ultimately obtaining the correct answer, while PAV models step-level progress under a prover policy (Zhang et al., 2026; Setlur et al., 2025). PURE instead targets the aggregation mechanism itself, showing that summing future step rewards can allow a few highly rewarded but unproductive steps to mask erroneous reasoning (Cheng et al., 2025). These methods are therefore particularly relevant to process reward models and multi-step reasoning, rather than to generic response-level reward hacking.

6. **Long-horizon and inference-time exploitation.** Some failures are created not by an inaccurate reward model alone but by the optimization procedure exploiting opportunities that emerge over multiple steps. MONA directly targets such multi-step reward-hacking strategies by combining myopic optimization with non-myopic approval signals (Farquhar et al., 2025). At inference time, HedgeTune instead limits the amount of proxy-guided search in Best-of-N-type procedures (Khalaf et al., 2026). These approaches are therefore most appropriate when the principal source of hacking is the strength or horizon of the search, rather than a specific artifact in the reward model.

7. **Judge-specific and explanation-specific vulnerabilities.** Generative reward models introduce additional failure modes. Master-RMs target "master key" phrases that cause an LLM judge to assign high reward without corresponding reasoning (Zhao et al., 2026). In explanation evaluation, Ferreira et al. show that process reward models may reward plausible post-hoc explanations that are not causally responsible for the model's decision (Ferreira et al., 2025). These failures require methods that distinguish genuine reasoning or causal influence from superficially plausible text and therefore cannot be fully addressed by generic reward regularization.

5.2 What Exactly Do the Mitigation Methods Change?

The most important distinction among the reviewed methods is the component of the reward-optimization pipeline that they modify. Table 2 therefore provides a unified comparison of the reviewed mitigation methods according to the failure type they target, the component at which they intervene, and whether the intervention occurs during training or inference. This organization makes it easier to compare methods that address similar failure modes through different parts of the reward-optimization pipeline.

The table highlights three broad intervention points. First, *reward-model and evaluator interventions* attempt to make the proxy more faithful or reduce the influence of unreliable reward estimates. InfoRM, ODIN, RRM, PRISM, and CROME primarily target spurious features and shortcut learning, but differ in how the problematic correlations are identified or removed. CRA and RATE instead use causal analysis to identify or correct problematic reward dependencies. Reward-model ensembles and uncertainty-aware routing address uncertainty more directly, reducing the probability that an unreliable reward estimate determines the optimization outcome.

Second, *policy-optimization interventions* accept that the reward model may remain imperfect and instead constrain how aggressively the policy can exploit it. ORPO, EPPO, Gradient Regularization, Constrained RLHF, RPO, BSPO, and ADVPO differ in the signal or constraint they use, but share the goal of limiting optimization in regions where the proxy becomes unreliable. These methods therefore intervene primarily on the policy or optimization procedure rather than attempting to eliminate every imperfection in the reward model.

Third, several methods intervene at *specialized stages of the alignment pipeline*. CRM, PAV, PURE, and causal attribution methods focus on process-level reasoning, credit assignment, or evaluation, while HedgeTune addresses inference-time search and MONA addresses multi-step exploitation. These methods illustrate why reward hacking should not be treated as a single failure mechanism: the appropriate intervention depends on whether the dominant problem lies in the reward representation, reward evaluation, policy optimization, temporal credit assignment, or inference-time search.

Within the reward-model family, the methods also differ substantially in scope. ODIN is particularly specialized to length-related reward hacking, whereas InfoRM, RRM, PRISM, and CROME address broader

Method	Failure type	Intervention	Stage
InfoRM	Spurious features and reward-model misgeneralization	Filters irrelevant information through a variational information bottleneck; uses latent-space outlier detection for overoptimized responses	Training
ODIN	Length-related reward hacking	Separates length-correlated information from reward-relevant representations using a two-head architecture and orthogonality constraint	Training
RRM	Prompt-independent artifacts	Uses causal augmentation to disrupt correlations between artifacts and preference labels	Training
PRISM	General shortcut learning	Enforces reward invariance across predefined shortcut groups rather than targeting a single artifact	Training
CROME	Unknown or diverse spurious attributes	Uses causal and neutral synthetic augmentations to distinguish genuine quality attributes from spurious ones	Training
CRA	Latent semantic confounding	Applies post-hoc causal backdoor adjustment using identified confounding features	Inference
RATE	Unknown causal influence of response attributes	Uses counterfactual rewriting to estimate causal effects of response attributes and audit reward predictions	Training / Analysis
Reward model ensembles	Reward-estimation variability	Aggregates predictions from multiple independently trained reward models	Inference
Uncertainty-aware routing	Unreliable or OOD reward predictions	Routes uncertain samples to a stronger judge while using a cheaper reward model for confident samples	Inference
Master-RMs	Superficial "master key" phrases	Trains generative reward models on adversarially truncated responses to reduce sensitivity to superficial reasoning cues	Training
CRM	Ambiguous step-level credit assignment	Defines step rewards through the conditional probability of ultimately reaching the correct answer	Training
PAV	Poor estimation of step-level progress	Predicts process advantage under a distinct prover policy to improve exploration	Training
PURE	Bad reasoning masked by reward aggregation	Replaces cumulative reward summation with minimum-based trajectory aggregation	Training
Causal attribution for explanations	Plausible but fabricated explanations	Adds causal attribution signals to distinguish genuine decision evidence from post-hoc rationalizations	Training
ORPO	Overoptimization of correlated proxies	Regularizes the policy occupancy measure toward a reference policy using χ^2 divergence	Training
EPPO	Reward hacking associated with internal dynamics	Penalizes deviations in energy loss from a supervised fine-tuning reference	Training
Gradient Regularization	Exploitation of sharp reward maxima	Penalizes large policy gradients and biases optimization toward flatter regions	Training
Constrained RLHF	Overoptimization of individual reward components	Constrains each reward component using its estimated proxy point rather than fixed scalar weights	Training
RPO	Distribution shift and OOD reward uncertainty	Combines DPO with an SFT regularizer to limit deviation from the preference-data distribution	Training
BSPO	Reward overestimation on unsupported actions	Restricts policy iteration using a behavior-supported Bellman operator	Training
ADVPO	Optimization against uncertain rewards	Optimizes against pessimistic reward functions within an uncertainty set estimated from reward-model representations	Training
HedgeTune	Inference-time overoptimization	Limits Best-of-N optimization by selecting the degree of search using limited true-reward calibration	Inference
MONA	Multi-step reward-hacking strategies	Uses myopic optimization with non-myopic approval signals to reduce incentives for long-horizon hacking	Training

Table 2: Unified comparison of reward-hacking mitigation methods. Methods are compared by the primary failure type they target, where they intervene in the reward-optimization pipeline, and whether the intervention occurs during training or inference.

classes of spurious features and shortcut learning. InfoRM compresses irrelevant information and uses the resulting representation for detection; RRM disrupts known artifact correlations through causal augmentation; PRISM imposes an explicit invariance structure; and CROME attempts to distinguish causal from spurious attributes without requiring the shortcut types to be specified beforehand.

CRA occupies a different position because it does not require changing the reward model itself. Instead, it applies causal adjustment after reward-model training, making it attractive when retraining the reward

model or policy is undesirable. Its limitation is that the relevant latent confounders must be identifiable well enough for the adjustment to be meaningful. RATE is also distinct because its primary contribution is diagnostic: by estimating the causal effects of individual attributes, it can reveal why a reward model behaves as it does, but causal diagnosis alone does not guarantee that subsequent policy optimization will be safe.

The process-reward methods likewise operate at different points in the reasoning pipeline. CRM changes the definition of the step-level reward, PAV changes what the verifier is trained to predict, PURE changes how step rewards are aggregated, and causal attribution methods change what constitutes evidence for a faithful explanation. Because these interventions address different components of process evaluation, they can in principle be combined rather than being strictly competing alternatives.

Overall, the unified comparison shows that the distinction between reward-model improvement and optimization constraints is useful but not absolute. Some methods improve the proxy itself, others modify how it is evaluated, and others constrain how strongly a policy or inference procedure can optimize it. Consequently, the most informative way to compare mitigation methods is to consider both *where they intervene in the reward-optimization pipeline* and *which failure mechanism they are designed to prevent*.

5.3 Overall Trade-off

The central trade-off across the literature is therefore between *reward fidelity*, *optimization freedom*, and *computational cost*. Methods that modify the reward model attempt to increase fidelity but may require new training data, architectural changes, or assumptions about shortcut structure. Methods that constrain optimization can remain effective with an imperfect reward model, but excessive conservatism may prevent the policy from discovering genuinely high-quality solutions. Uncertainty-aware and routing approaches can preserve performance while concentrating computation on difficult cases, but they require reliable uncertainty estimation and, in some cases, access to a substantially stronger judge.

More fundamentally, the literature indicates that reward hacking cannot be reduced to a single defect of reward models. The failure can arise from *what the reward model represents*, *where the policy operates*, *how strongly the proxy is optimized*, *how rewards are assigned across reasoning steps*, or *how much search is performed at inference time*. Consequently, the appropriate mitigation should be selected according to the observed failure mechanism rather than according to a generic notion of "reward-model robustness." A reward-model intervention is most direct when the source of reward error is identifiable; an optimization constraint is more appropriate when the reward is useful but becomes unreliable under strong optimization; uncertainty-based methods are suitable when reliability varies across samples; and specialized PRM, judge, or inference-time methods are necessary when the hacking mechanism is specific to reasoning, generative evaluation, or search.

Overall, the strongest implication of the comparison is that improving reward models and constraining optimization should be viewed as complementary rather than competing strategies. Reward-model improvements reduce the set of exploitable weaknesses, while conservative or uncertainty-aware optimization reduces the opportunity to exploit weaknesses that remain. Nevertheless, the theoretical results on proxy alignment imply that neither strategy can provide a general guarantee of unhackability. The practical objective is therefore not to eliminate the possibility of reward hacking altogether, but to identify the dominant failure mechanism and intervene at the corresponding level of the reward-optimization pipeline.

6 Open Problems and Future Directions

Unified definitions and evaluation protocols. Existing work provides complementary but incompatible formalisms of reward hacking, ranging from MDP-based definitions to correlation-based and empirical scaling perspectives. A unified evaluation framework applicable across RLHF, direct preference optimization, and inference-time selection remains an open problem.

Detection of overoptimization. Several studies propose internal signals for detecting reward hacking, including latent-space structure and energy-based measures. However, robust and general detection methods that transfer across models, tasks, and training regimes remain underdeveloped.

Scaling behavior of mitigation strategies. While larger reward models and conservative optimization both reduce overoptimization to some extent, their scaling behavior is not fully understood. In particular, whether mitigation techniques remain effective as model and dataset scales increase is still unclear.

Out-of-distribution generalization for reward models. RRM and GAS-DRO highlight the importance of improving reward model robustness beyond the training preference distribution (Liu et al., 2025; Wen & Yang, 2026). A key future direction is to develop reward models that generalize across distribution shifts introduced by evolving policies, diverse response styles, and previously unseen optimization behaviors. Robust reward models should capture stable human-aligned signals that transfer to novel settings and remain reliable against emerging reward-hacking strategies.

Evaluation Beyond Preference Accuracy. Future work should develop evaluation protocols that explicitly measure a reward model's robustness to optimization rather than relying solely on static preference prediction benchmarks. Since reward models are ultimately optimized by downstream algorithms such as RLHF and best-of-n sampling, evaluation should account for the extent to which optimization induces reward overoptimization. Incorporating optimization-aware metrics into reward-model assessment could help identify reward models that remain reliable under increasing optimization pressure and reduce the risk of reward hacking in aligned language models (Kim et al., 2025).

7 Conclusion

This survey reviewed recent work on reward hacking in large language model alignment, spanning theoretical characterizations, empirical analyses, and mitigation strategies. Across these settings, a consistent pattern emerges: optimization of imperfect proxy objectives can lead to systematic divergence between proxy performance and true task quality.

Existing mitigation approaches can be broadly categorized into methods that improve reward model robustness and methods that constrain optimization dynamics. Although both categories reduce specific failure modes, neither fully resolves reward hacking under strong optimization regimes. The reviewed literature therefore suggests that robust alignment requires a combination of improved proxy design and conservative optimization procedures.

Acknowledgments

We thank the course instructors for their guidance throughout this project and for assembling the paper corpus used in this survey.

References

Johannes Ackermann, Michael Noukhovitch, Takashi Ishida, and Masashi Sugiyama. Gradient regularization prevents reward hacking in reinforcement learning from human feedback and verifiable rewards. *arXiv preprint arXiv:2602.12345*, 2026.

Lichang Chen, Chen Zhu, Davit Soselia, Jiuhai Chen, Tianyi Zhou, Tom Goldstein, Heng Huang, Mohammad Shoeybi, and Bryan Catanzaro. Odin: Disentangled reward mitigates hacking in rlhf. *arXiv preprint arXiv:2402.07319*, 2024.

Jie Cheng, Gang Xiong, Ruixi Qiao, Lijun Li, Chao Guo, Junle Wang, Yisheng Lv, and Fei-Yue Wang. Stop summation: Min-form credit assignment is all process reward model needs for reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Thomas Coste, Usman Anwar, Robert Kirk, and David Krueger. Reward model ensembles help mitigate overoptimization. In *International Conference on Learning Representations*, 2024.

Juntao Dai, Taiye Chen, Yaodong Yang, Qian Zheng, and Gang Pan. Mitigating reward over-optimization in rlhf via behavior-supported regularization. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv preprint arXiv:2503.18130.

Sebastian Farquhar, Vikrant Varma, David Lindner, David Elson, Caleb Biddulph, Ian Goodfellow, and Rohin Shah. Mona: Myopic optimization with non-myopic approval can mitigate multi-step reward hacking. *arXiv preprint arXiv:2501.13011*, 2025.

Pedro Ferreira, Wilker Aziz, and Ivan Titov. Truthful or fabricated? using causal attribution to mitigate reward hacking in explanations. *arXiv preprint arXiv:2504.05294*, 2025.

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

Hadi Khalaf, Claudio Mayrink Verdun, Alex Oesterling, Himabindu Lakkaraju, and Flavio Calmon. Inference-time reward hacking in large language models. *Advances in Neural Information Processing Systems*, 38:61720–61760, 2026.

Sunghwan Kim, Dongjin Kang, Taeyoon Kwon, Hyungjoo Chae, Dongha Lee, and Jinyoung Yeo. Rethinking reward model evaluation through the lens of reward overoptimization. *arXiv preprint arXiv:2505.12763*, 2025.

Cassidy Laidlaw, Shivam Singhal, and Anca Dragan. Correlated proxies: A new definition and improved mitigation for reward hacking. In *International Conference on Learning Representations*, 2025.

Tianqi Liu, Wei Xiong, Jie Ren, Lichang Chen, Junru Wu, Rishabh Joshi, Yang Gao, Jiaming Shen, Zhen Qin, Tianhe Yu, Daniel Sohn, Anastasiia Makarova, Jeremiah Liu, Yuan Liu, Bilal Piot, Abe Ittycheriah, Aviral Kumar, and Mohammad Saleh. RRM: Robust reward model training mitigates reward hacking. In *International Conference on Learning Representations*, 2025.

Zhihan Liu, Miao Lu, Shenao Zhang, Boyi Liu, Hongyi Guo, Yingxiang Yang, Jose Blanchet, and Zhaoran Wang. Provably mitigating overoptimization in rlhf: Your sft loss is implicitly an adversarial regularizer. *Advances in Neural Information Processing Systems*, 37:138663–138697, 2024.

Yuchun Miao, Sen Zhang, Liang Ding, Rong Bao, Lefei Zhang, and Dacheng Tao. Inform: Mitigating reward hacking in rlhf via information-theoretic reward modeling. *Advances in Neural Information Processing Systems*, 37:134387–134429, 2024.

Yuchun Miao, Sen Zhang, Liang Ding, Yuqi Zhang, Lefei Zhang, and Dacheng Tao. The energy loss phenomenon in rlhf: A new perspective on mitigating reward hacking. *arXiv preprint arXiv:2501.19358*, 2025.

Ted Moskovitz, Aaditya Singh, DJ Strouse, Tuomas Sandholm, Ruslan Salakhutdinov, Anca Dragan, and Stephen McAleer. Confronting reward model overoptimization with constrained rlhf. In *International Conference on Learning Representations*, volume 2024, pp. 21998–22025, 2024.

Rafael Rafailov, Yaswanth Chittepu, Ryan Park, Harshit Sikchi, Joey Hejna, W. Bradley Knox, Chelsea Finn, and Scott Niekum. Scaling laws for reward model overoptimization in direct alignment algorithms. In *Advances in Neural Information Processing Systems*, 2024.

David Reber, Sean Richardson, Todd Nief, Cristina Garbacea, and Victor Veitch. Rate: Causal explainability of reward models with imperfect counterfactuals. *arXiv preprint arXiv:2410.11348*, 2024.

Amrith Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. Rewarding progress: Scaling automated process verifiers for LLM reasoning. In *International Conference on Learning Representations (ICLR)*, 2025.

Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward hacking. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Ruike Song, Zeen Song, Huijie Guo, and Wenwen Qiang. Causal reward adjustment: Mitigating reward hacking in external reasoning via backdoor correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 33019–33027, 2026.

Pragya Srivastava, Harman Singh, Rahul Madhavan, Gandharv Patil, Sravanti Addepalli, Arun Suggala, Rengarajan Aravamudhan, Soumya Sharma, Anirban Laha, Aravindan Raghuveer, et al. Robust reward modeling via causal rubrics. *arXiv preprint arXiv:2506.16507*, 2025.

Jiaqi Wen and Jianyi Yang. Distributionally robust optimization via generative ambiguity modeling. In *International Conference on Learning Representations*, 2026.

Jiaxin Wen, Ruiqi Zhong, Akbir Khan, Ethan Perez, Jacob Steinhardt, Minlie Huang, Sam Bowman, He He, and Shi Feng. Language models learn to mislead humans via rlhf. In *International Conference on Learning Representations*, volume 2025, pp. 74670–74692, 2025.

Zaiyan Xu, Sushil Vemuri, Kishan Panaganti, Dileep Kalathil, Rahul Jain, and Deepak Ramachandran. Robust llm alignment via distributionally robust direct preference optimization. *Advances in Neural Information Processing Systems*, 38:27039–27076, 2026.

Zhenghao Xu, Qin Lu, Qingru Zhang, Liang Qiu, Ilgee Hong, Changlong Yu, Wenlin Yao, Yao Liu, Haoming Jiang, Lihong Li, Hyokun Yun, and Tuo Zhao. Ask a strong LLM judge when your reward model is uncertain. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Wenqian Ye, Guangtao Zheng, and Aidong Zhang. Rectifying shortcut behaviors in preference-based reward learning. *Advances in Neural Information Processing Systems*, 38:64712–64740, 2026.

Xiaoying Zhang, Jean-François Ton, Wei Shen, Hongning Wang, and Yang Liu. Mitigating reward overoptimization via lightweight uncertainty estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Zheng Zhang, Ziwei Shan, Kaitao Song, Yexin Li, and Kan Ren. Linking process to outcome: Conditional reward modeling for LLM reasoning. In *International Conference on Learning Representations (ICLR)*, 2026. URL `https://foundation-model-research.github.io/CRM`.

Yulai Zhao, Haolin Liu, Dian Yu, S.Y. Kung, Meijia Chen, Haitao Mi, and Dong Yu. One token to fool LLM-as-a-judge. *arXiv preprint arXiv:2502.12345*, 2026.