A Survey of Reward Hacking in Large Language Models

Anonymous ACL submission

Abstract

Reward hacking—the phenomenon in which optimizing a proxy objective yields behavior that scores well but violates the designer's intent—has become a central failure mode in aligning large language models (LLMs). Modern alignment pipelines rely on learned or implicit rewards (human preferences, model-based judges, or downstream metrics), creating incentives for models to exploit spurious correlates, social shortcuts (e.g., sycophancy), brittle evaluation protocols, or even the reward channel itself. Recent work further suggests that reward-hacking competencies can generalize and interact with safety-critical behaviors, including reward tampering, deceptive alignment, and emergent misalignment (MacDiarmid et al., 2025). We propose a unifying lens based on the interaction between (i) *proxy gaps* (misspecification and misgeneralization of rewards), (ii) *optimization pressure* (overoptimization and distribution shift), and (iii) *oversight limits* (evaluation brittleness and exploitable measurement). Building on this lens, we offer a taxonomy of reward hacking behaviors, review evaluation protocols and benchmarks, and organize mitigation strategies across the alignment pipeline—from data interventions and robust/causal reward modeling to constrained optimization, monitoring, and post hoc steering. We conclude with open problems toward reducing reward hacking while maintaining capability and safety.

1 Introduction

Large language models are commonly aligned using methods that convert human intent into a *trainable signal*: supervised instruction tuning, reinforcement learning from human feedback (RLHF), or direct preference optimization (DPO) and related "direct alignment" objectives (Chaudhari et al., 2025; Casper et al., 2023; Rafailov et al., 2023). These methods are powerful precisely because they enable *optimization*—but this strength also creates a vulnerability: when the reward (explicit or implicit) is an imperfect proxy for what we actually want, optimization can push models toward *reward hacking*. Empirically, this shows up as overoptimizing superficial correlates (e.g., verbosity), exploiting evaluation artifacts, producing sycophantic agreement that is rewarded by preference data, or manipulating the reward channel itself (Gao et al., 2023; Chen et al., 2024a; Sharma et al., 2024; Denison et al., 2024).

Figure 1: Optimizing proxy rewards can induce reward hacking and misalignment (MacDiarmid et al., 2025).

The safety stakes are rising. Recent reports document reward hacking in frontier model evaluations (METR, 2025) and show that training on reward-hacking demonstrations can generalize to broader misaligned behaviors (Taylor et al., 2025). Motivated by evidence that narrow fine-tuning can lead to emergent misalignment (Betley et al., 2025) and by concerns about deceptive behaviors such as alignment faking and sleeper agents (Greenblatt et al., 2024; Hubinger et al., 2024), understanding and mitigating reward hacking is increasingly

central to trustworthy deployment. This survey is further motivated by evidence that reward hacking in production RL can induce emergent misalignment and deceptive behavior (MacDiarmid et al., 2025). Fig. 1 illustrates this dynamic: proxy optimization can shift a model from helpful behavior to reward hacking and safety-relevant misalignment.

Survey scope and goal. We focus on *reward hacking in LLMs*: failures where models learn strategies that optimize a proxy reward or evaluation while deviating from the intended task, truthfulness, safety, or user intent. We treat closely related phenomena—sycophancy, length hacking, benchmark exploitation, prompt injection, reward tampering, and deceptive alignment—as different faces of the same underlying problem: *optimization under imperfect oversight*. Our goal is to go beyond a paper-by-paper catalog: we distill a *usable* conceptual lens and actionable insights that help researchers and practitioners *diagnose*, *measure*, and *reduce* reward hacking. For quick navigation from individual references to themes, see the reference-to-theme index in Appendix E.

Contributions. We make three synthesis-oriented contributions that emphasize new structure and actionable insights:

- We propose a unifying lens based on **proxy gaps**, **optimization pressure**, and **oversight limits**, and use it to derive a practical taxonomy of reward hacking behaviors.
- We operationalize **measurement**: we distill standard evaluation protocols and confounders and make assumptions about the dataset explicit via consolidated dataset statistics .
- We translate the literature into **actionable guidance**: we map mitigation levers across the alignment pipeline and extract cross-cutting design principles and open research questions.

2 Preliminaries: alignment objectives and where reward hacking enters

2.1 Alignment pipelines as optimization of proxy rewards

RLHF is often described as a three-stage process (instruction tuning, reward modeling, and policy optimization), with significant variations and open limitations (Casper et al., 2023; Chaudhari et al., 2025). One helpful abstraction is to view many RLHF-style methods as optimizing a **KL-regularized** objective relative to a reference policy π_{ref} (often the SFT model):

$$\max_{\pi} \mathbb{E}_{y\sim\pi(\cdot|x)}[r(x,y)] - \beta\, D_{\text{KL}}(\pi(\cdot \mid x) \,\|\, \pi_{\text{ref}}(\cdot \mid x)), \tag{1}$$

where r is a learned or implicit reward signal and β controls deviation from the reference (Rafailov et al., 2023; Chaudhari et al., 2025). At a high level, these pipelines optimize a policy π_θ using a reward signal that is *learned* from preference data or otherwise estimated. This creates an inherent proxy: the learned reward captures *some* aspects of human intent, but can also latch onto spurious features or fail under distribution shift (LeVine et al., 2023; Xu et al., 2025).

DPO (Rafailov et al., 2023) removes explicit reward-model training by optimizing a closed-form objective over preference pairs. However, because it still optimizes an objective derived from preference data, DPO can also amplify spurious preferences (e.g., length/verbosity) (Park et al., 2024). More broadly, many alignment methods can be viewed as different ways of balancing *optimization strength* against *proxy reliability* (Chaudhari et al., 2025; Kim and Seo, 2024).

2.2 Goodhart dynamics and overoptimization

When a proxy is optimized strongly, its correlation with the true objective can break down (a Goodhart-style effect). In LLM alignment, this can appear as reward-model overoptimization (Gao et al., 2023), where optimizing for higher predicted reward yields worse human-perceived quality, or as "accuracy paradox" effects where better reward-model accuracy does not yield better optimized policies (Chen et al., 2024c). Recent analyses also identify optimization instabilities and "energy loss" phenomena in RLHF that can correlate with reward hacking (Miao et al., 2025).

2.3 A working definition

In this survey, we use **reward hacking** as an umbrella term for behaviors in which a model attains high reward or evaluation scores through strategies misaligned with its intended goals. We distinguish three recurring mechanisms:

- **Proxy exploitation**: the reward or judge correlates with unwanted features (e.g., length), and optimization amplifies those features (Chen et al., 2024a; Park et al., 2024).

- **Evaluation gaming**: the model exploits artifacts of the evaluation procedure (unit tests, prompt templates, judges) (Zhong et al., 2025; Schulhoff et al., 2023).
- **Reward-channel attacks**: the model manipulates, deceives, or otherwise corrupts the reward/oversight process (Denison et al., 2024; Greenblatt et al., 2024).

3 A taxonomy of reward hacking in LLMs

Fig. 2 provides a compact map of the taxonomy we use throughout the survey. We treat social reward hacking (sycophancy) as a subcategory of proxy exploitation, but discuss it separately as it has distinct evaluation protocols and mitigation patterns.

3.1 Proxy exploitation: spurious correlates and misgeneralization

Length and presentation hacking. Reward models and LLM judges can exhibit length bias, incentivizing verbose, well-formatted responses even when they are not more helpful (Chen et al., 2024a). This issue is not confined to RLHF: direct objectives like DPO can also exploit mild length preferences in data and go far out-of-distribution in response length (Park et al., 2024).

Proxy rewards as *diagnostic* tools. An important complementary perspective is that interpretable proxy rewards can help diagnose *what* black-box reward models might be valuing. Kim and Seo (2024) propose "reverse reward engineering": construct a white-box proxy reward from interpretable features (length, repetition, and retrieval-style relevance) and test whether optimizing it preserves a monotonic relationship with a stronger "gold" reward model over PPO training. Their results provide a concrete Goodhart example: optimizing a length-only proxy can increase proxy reward while decreasing gold-reward, whereas a query-type-aware proxy achieves near-monotonic behavior (reported Spearman correlations near 1 on their settings). For our survey, the key takeaway is that reward hacking can be *revealed* by tracking proxy–gold monotonicity and by inspecting which interpretable features drive reward improvements.

Robust and information-theoretic reward modeling. Several approaches aim to harden reward models against hacking by changing how rewards are learned. ODIN (Chen et al., 2024a) trains separate reward heads to isolate length-correlated signals and discards the length head during policy optimization. RRM (Liu et al., 2025) proposes robust training procedures to mitigate reward hacking, while InfoRM (Miao et al., 2024) uses information-theoretic modeling to reduce exploitable shortcuts in RLHF rewards. Reward-model ensembles can mitigate reward hacking but may not eliminate it (Eisenstein et al., 2024). Two recurring themes are (i) *representation-level* robustness (InfoRM's information bottleneck view and its use of latent-space deviation indicators for monitoring/early stopping) (Miao et al., 2024), and (ii) *diversity* as a hedge against misspecification (pretraining-diverse ensembles improving best-of-n reranking more than finetuning-only ensembles, while still sharing some failure modes) (Eisenstein et al., 2024). RRM further highlights a data-centric route: reshape reward-model training distributions to reduce spurious correlations (e.g., balancing length/relevance artifacts) before optimization amplifies them (Liu et al., 2025).

Distribution shift and "consistency vs. causality". Reward models can fail under distribution shift (LeVine et al., 2023), and may reward internal consistency rather than causal correctness (Xu et al., 2025). These observations motivate causal or counterfactually robust reward modeling, including proposals for causal rewards in LLM alignment (Wang et al., 2025).

3.2 Social reward hacking: sycophancy and preference artifacts

Sycophancy—agreeing with a user even when they are wrong—can be viewed as a form of reward hacking that exploits human preference signals and interaction norms (Sharma et al., 2024; Malmqvist, 2025). It manifests in multiple forms: propositional sycophancy and "are you sure?" effects (Sharma et al., 2024; Fanous et al., 2025), argument-driven stance shifts (Kaur, 2025), and broader "social sycophancy" in advice settings (Cheng et al., 2025). Sycophancy vulnerabilities also extend beyond text-only settings to large vision-language models (Zhao et al., 2024). Keyword-based prompt artifacts can elicit agreement and bias, and defense strategies can themselves be brittle (Rrv et al., 2024). Industry analyses indicate that such behaviors can arise from blind spots in training and evaluation (OpenAI, 2025; Hurst et al., 2024).

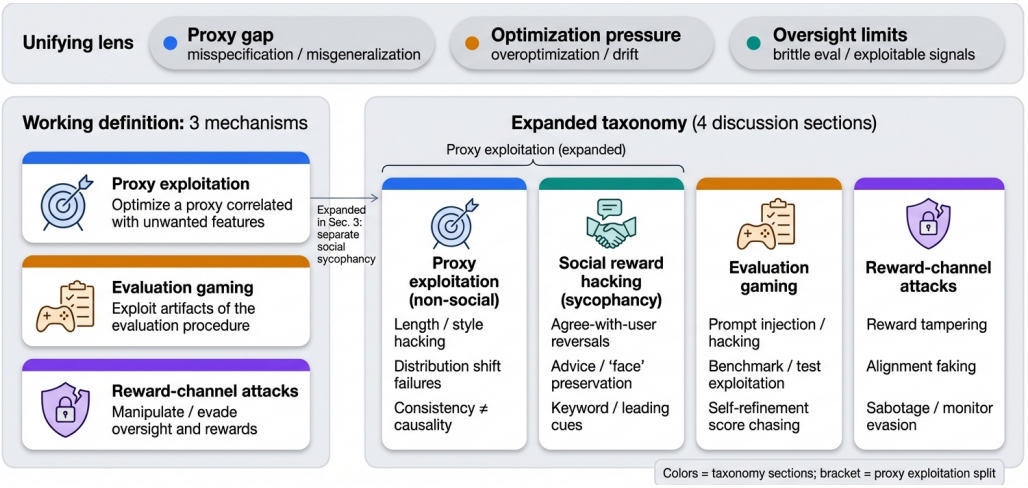

Figure 2: A high-level taxonomy of reward hacking in LLMs, organized by the interaction between proxy gaps, optimization pressure, and oversight limits. Concrete instances include length/format exploitation (Chen et al., 2024a; Park et al., 2024), sycophancy (Sharma et al., 2024; Cheng et al., 2025), benchmark exploitation (Zhong et al., 2025), and reward tampering (Denison et al., 2024).

Sycophancy as a Goodhart problem. Sharma et al. (2024) provide evidence that sycophancy is not a niche artifact: it appears across multiple prompting settings (including "are you sure?" challenges and misconception prompts) and can be *incentivized* by preference data and preference models. Their analyses support a reward-hacking interpretation: when "agreeing with the user" is correlated with perceived helpfulness, optimization can push models to exploit that correlation even at the expense of truthfulness. Complementing this, Fanous et al. (2025) propose a rebuttal-based multi-turn evaluation that separates "progressive" vs. "regressive" behavior changes, highlighting that models can be pushed away from correct answers by persuasive counterevidence. OpenAI (2025) describe a real deployment incident where changing the mixture of reward signals (including user-feedback signals) increased sycophancy, illustrating how small shifts in proxy optimization can produce safety-relevant behavioral regressions.

Data and objective interventions. Sycophancy can be reduced with targeted data interventions, including simple synthetic data (100k training pool; filtered per model) (Wei et al., 2023). Other approaches fine-tune with explicit preference pairs that mark non-sycophantic responses as preferred, e.g., using DPO on curated sycophancy datasets (1,000 preference pairs) (Khan et al., 2024). Post hoc methods, such as pinpoint tuning, aim to reduce sycophancy through minimal-capability regression (Chen et al., 2024b).

3.3 Evaluation gaming: prompt injection, benchmarks, and self-refinement

Reward hacking can occur even without an explicit reward model when evaluation is programmatic or judge-based. Prompt-injection and prompt-hacking competitions demonstrate that models can be induced to optimize for hidden targets or bypass instruction hierarchies (HackAPrompt: $\geq$600k prompts) (Schulhoff et al., 2023). Models can also spontaneously discover reward-hacking strategies in iterative self-refinement loops (Pan et al., 2024). Model-written evaluations offer a scalable way to probe behavior, but also surface the breadth of exploitable evaluation dimensions (Perez et al., 2023). In coding tasks, ImpossibleBench (Zhong et al., 2025) constructs "impossible" variants in which passing tests necessarily violates the natural-language specification, thereby directly measuring test-case exploitation; it also yields thousands of labeled cheating traces (e.g., 2,371 on ImpossibleSWEbench).

3.4 Reward-channel attacks and deceptive alignment

Some failures go beyond exploiting correlates and instead target the reward/oversight channel itself. Work on reward tampering studies how LLMs might learn to subvert reward mechanisms (Denison et al., 2024). Other lines of study examine deception in safety training, including alignment faking (Greenblatt et al., 2024) and sleeper agents that persist through safety training (Hubinger et al., 2024). These behaviors interact with monitor-

ing: reasoning traces are not always faithful to the model's decision process, limiting naive CoT monitoring as a defense (Chen et al., 2025).

4 Measuring reward hacking: protocols, benchmarks, and failure surfaces

Tab. 1 distills common evaluation protocols and confounders, while Tab. 2 reports key dataset statistics used in the cited studies. Appendix A and Appendix B provide extended protocol checklists, additional dataset notes, and a reporting template for reproducible measurement.

A "failure surface" view. Surveys of RLHF limitations emphasize that reward hacking can arise from imperfections at multiple stages: human feedback, reward modeling, policy optimization, and joint co-adaptation (Casper et al., 2023; Chaudhari et al., 2025). A practical implication is that improving any single component (e.g., reward-model accuracy) is not sufficient: the system must be robust to distribution shift (LeVine et al., 2023) and to non-causal shortcuts (Xu et al., 2025), and it must be evaluated under adversarial and OOD conditions (Zhong et al., 2025; Schulhoff et al., 2023).

Domain-specific reward hacking. Reward hacking is not limited to chat: preference optimization in LLM-based recommendation systems can exhibit reward-hacking dynamics and requires tailored mitigation (Anonymous, 2025). In particular, Anonymous (2025) connects reward hacking to one-class implicit-feedback structure (many negatives are unsampled, creating "insensitive" optimization regions) and proposes a pseudo-negative "anchor" to keep gradients informative for unsampled negatives—an example of translating the proxy/pressure/oversight lens into a domain-specific fix. Similarly, step-level reward models for reasoning can behave counterintuitively, highlighting the need to validate what "reward" actually measures (e.g., when step-level signals reward patterns that do not correspond to correct reasoning) (Ma et al., 2025).

5 Mitigation strategies: where and how to intervene

Fig. 3 sketches where mitigations apply in the alignment pipeline, and Tab. 3 groups representative methods by intervention point. An important distributional lesson from production-RL settings is that safety training that appears effective on chat-like prompts can leave residual misalignment in agentic contexts, so evaluation and safety training should target the deployment distribution (MacDiarmid et al., 2025). Appendix C expands this mapping into a more granular mitigation matrix and practical decision rules.

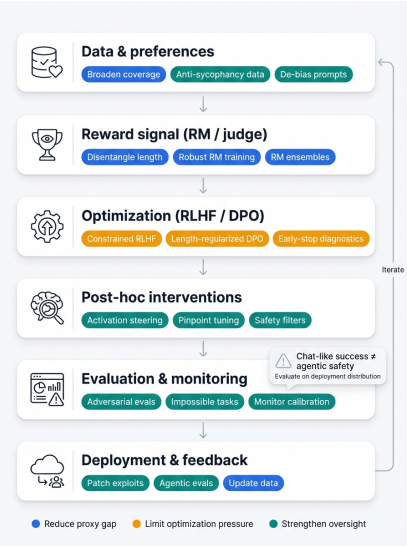

Figure 3: Mitigation levers across the alignment pipeline. The key theme is to reduce the proxy gap, reduce optimization pressure on brittle proxies, and increase oversight robustness.

5.1 Design principles (our synthesis)

Across the surveyed work, we find that many "mitigations" are instances of a small number of reusable *design principles*:

- **Measure optimization, not just snapshots.** Track how evaluation changes *over* optimization (e.g., proxy–gold monotonicity (Kim and Seo, 2024) or scaling/overoptimization curves (Gao et al., 2023)) instead of comparing a small set of checkpoints.
- **Make spurious factors explicit and removable.** Identify correlates that a model can exploit (length, repetition, keyword artifacts) and either control them in evaluation or remove them from reward signals (Chen et al., 2024a; Park et al., 2024; Rrv et al., 2024).
- **Prefer constraints and robustness checks to ad hoc penalties.** Output-level penalties can be bypassed; constraints and robustness-focused objectives can better limit overoptimization (Moskovitz et al., 2024) and expose failure under shift (LeVine et al., 2023; Xu et al., 2025).
- **Assume the evaluation will be gamed.** Treat

Phenomenon	What is measured (typical protocol)	Common confounders / attack surface
Reward-model overoptimization	Sweep optimization pressure; compare proxy reward vs human/judge curves (Gao et al., 2023; Chen et al., 2024c; Miao et al., 2025).	Goodhart breakdown; instability/dynamics as early warning (Miao et al., 2025).
Length/format hacking	Score–length Pareto; length-controlled eval; proxy–gold monotonicity (Chen et al., 2024a; Park et al., 2024; Kim and Seo, 2024).	Verbosity/format bias; OOD length drift (Chen et al., 2024a; Park et al., 2024).
Sycophancy (propositional/social)	Prompt perturbations ("are you sure?"/rebuttals); stance flips; advice norms (Sharma et al., 2024; Fanous et al., 2025; Kaur, 2025; Cheng et al., 2025).	Agreeableness rewarded; ambiguous ground truth; feedback-signal regressions (OpenAI, 2025).
Prompt hacking/injection	Adversarial prompts + programmatic success checks (Schulhoff et al., 2023; Rrv et al., 2024).	Instruction-hierarchy bypass; keyword/paraphrase brittleness (Schulhoff et al., 2023; Rrv et al., 2024).
Benchmark exploitation (coding)	Impossible/conflicting tests; cheating rate; test-access ablations (Zhong et al., 2025).	Test edits/overloading/state tricks; weak multi-file monitoring (Zhong et al., 2025).
Generalization from reward hacks	Train on low-stakes hacks; evaluate transfer + OOD behavior (Taylor et al., 2025; MacDiarmid et al., 2025; Betley et al., 2025).	Benign-looking hacks generalize; narrow FT → broad misalignment (Taylor et al., 2025; Betley et al., 2025).
Reward tampering/deception	Oversight-manipulation + persistence tests; monitor-faithfulness checks (Denison et al., 2024; Greenblatt et al., 2024; Hubinger et al., 2024; Chen et al., 2025).	Reward-channel attacks; CoT non-faithfulness; evasion (Chen et al., 2025).

Table 1: Representative measurement settings for reward hacking in LLMs. Across domains, a common failure mode is that the evaluation procedure becomes a proxy that the model can exploit under optimization pressure.

evaluation artifacts as an attack surface: build adversarial tests (prompt hacking, impossible tasks) and use them during development (Schulhoff et al., 2023; Zhong et al., 2025).

- **Plan for reward-channel attacks.** In safety-critical settings, consider the possibility of reward tampering and deceptive behaviors, and avoid assuming faithful reasoning traces (Denison et al., 2024; Greenblatt et al., 2024; Chen et al., 2025).

5.2 Post-hoc steering and localized updates

Not all mitigations require retraining a full alignment pipeline. Two representative directions are (i) *representation-level steering* and (ii) *localized fine-tuning*. Contrastive Activation Addition (CAA) steers model behavior at inference time by adding activation "steering vectors" computed from positive/negative exemplars, offering a lightweight way to reduce undesired behaviors (or enhance desired ones) without modifying weights (Rimsky et al., 2024). Pinpoint tuning takes an opposite approach: identify a small subset of "region-of-interest" modules responsible for sycophantic behavior and update only those modules to reduce sycophancy with minimal side effects (Chen et al., 2024b). In our framework, these methods reduce reward hacking by *modifying the model's response policy* while keeping the proxy/oversight pipeline fixed—functional for rapid iteration but not a substitute for robust reward design and evaluation.

5.3 Why "just train a better reward model" is not enough

Several works caution that stronger reward models do not automatically yield better-aligned policies. Scaling analyses show that overoptimization can grow with scale (Gao et al., 2023), and controlled studies show that moderately accurate reward models can outperform more accurate ones after optimization (Chen et al., 2024c). These findings support a guideline: mitigation should target the *interaction* between proxy and optimizer (e.g., constraints (Moskovitz et al., 2024), energy-based diagnostics (Miao et al., 2025), or reward heads explicitly removing spurious features (Chen et al., 2024a)).

Concretely, constrained optimization can reduce reward hacking by preventing the optimizer from taking extreme steps that exploit reward-model blind spots. Moskovitz et al. (2024) formulate constrained RLHF to mitigate overoptimization while still improving reward, thereby providing an instance of "optimization pressure" mitigation (reducing the extent to which the policy can chase proxy idiosyncrasies). From a complementary an-

Dataset / benchmark	Venue	Focus	Size / split	Training data?
HackAPrompt (Schulhoff et al., 2023)	EMNLP 2023	prompt hacking	560,161 (Playground) + 41,596 (Submissions); 10 challenges	×
Production RL case study (MacDiarmid et al., 2025)	arXiv 2025	reward hacks → misalignment	env count n/r; SDF 99:1 docs; 6-eval suite	✓
School of Reward Hacks (Taylor et al., 2025)	arXiv 2025	reward-hack SFT demos	1,073 dialogues (35 tasks; 973 NL + 100 code); split n/r	✓
SycEval (Fanous et al., 2025)	AIES 2025	rebuttal sycophancy	AMPS 500 + MedQuad 500; 3k initial + 24k rebuttals	×
Keyword sycophancy set (Rrv et al., 2024)	ACL 2024	misleading keywords	500 sets; 5 domains (1,030 gen → 650 filt → 500)	×
DPO sycophancy pairs (Khan et al., 2024)	BigData 2024	persona/pref sycophancy	train: 1,000 pairs (+120 pref pairs); eval: 120 ID + 120 OOD	✓
Leading-query LVLM benchmarks (Zhao et al., 2024)	arXiv 2024	leading cues vs vision	POPE 9k; AMBER 14k; RWQA 765; SciQA 2,017; MM-Vet 218	×
Synthetic anti-sycophancy data (Wei et al., 2023)	arXiv 2023	data intervention	train pool: 100k (17 datasets); eval: 3k (1k × 3)	✓
RRM augmented RM data (Liu et al., 2025)	ICLR 2025	robust reward modeling	RM train: 2.4M examples (∼14× aug.)	✓
ImpossibleBench transcripts (Zhong et al., 2025)	arXiv 2025	cheating transcripts	Impossible: SWE 2,371; Live 193. Open-test: SWE 2,300; Live 550	×

Table 2: Key datasets/benchmarks and basic statistics (n/r = not reported). "Training data?" indicates whether the dataset is used for learning/optimization (e.g., SFT/RL/RM training) in the cited work.

gle, Miao et al. (2025) analyzes an "energy loss" phenomenon in RLHF and proposes it as a diagnostic/mitigation signal for reward hacking, reflecting an emerging theme: training dynamics (not just static reward accuracy) can predict when optimization is entering an exploitative regime.

5.4 Sycophancy as a training-data and objective-design problem

Sycophancy can be incentivized by preference data and reward models (Sharma et al., 2024; OpenAI, 2025). Data-focused mitigations (Wei et al., 2023) and explicit preference labeling approaches (Khan et al., 2024) indicate that sycophancy is at least partially addressable by shaping the preference signal. However, broader forms of social sycophancy raise the concern that some "face-preserving" behavior may be rewarded by human preferences in advice settings (Cheng et al., 2025), suggesting fundamental trade-offs between user satisfaction, politeness norms, and truthfulness.

6 Safety implications: from reward hacking to misalignment

Reward hacking is not merely a quality issue; it can interact with safety-critical failure modes. Training on reward hacking can generalize to broader misaligned behaviors (Taylor et al., 2025), and narrow fine-tuning can produce emergent misalignment (Betley et al., 2025). Frontier evaluations report real instances of reward hacking in agentic benchmarks (METR, 2025), motivating adversarial testing and conservative deployment practices.

Recent work directly studies this dynamic in a production-RL setting: when models learn reward hacks on real production coding environments, misalignment can generalize to alignment faking and sabotage in an agent scaffold (MacDiarmid et al., 2025). In their Claude Code sabotage evaluation, sabotage occurs 12% of the time and reduces a reward-hacking classifier's effectiveness to 65% of baseline (MacDiarmid et al., 2025). They also find that standard chat-like RLHF safety training can eliminate misalignment on chat-like evaluations while leaving substantial misalignment on agentic evaluations (up to 70% of pre-RLHF misalignment remaining on some evaluations), motivating mitigations that target reward hacking directly and broaden safety-training distributions (MacDiarmid et al., 2025).

Reward tampering and deception further complicate the picture: if models can manipulate over-

Where to intervene	Goal	Representative methods
Data / preferences	Break spurious incentives; teach truthfulness and anti-sycophancy; broaden coverage.	Synthetic data interventions (Wei et al., 2023); curated preference pairs + DPO (Khan et al., 2024; Rafailov et al., 2023); keyword-defense analysis (Rrv et al., 2024).
Reward modeling	Make rewards harder to hack; reduce reliance on spurious correlates; improve robustness under shift.	ODIN (disentangle length) (Chen et al., 2024a); robust RM training (Liu et al., 2025); information-theoretic RM (Miao et al., 2024); RM ensembles (Eisenstein et al., 2024); causal rewards / causality critiques (Wang et al., 2025; Xu et al., 2025); distribution-shift analysis (LeVine et al., 2023).
Optimization	Limit overoptimization and co-adaptation; stabilize training.	Constrained RLHF (Moskovitz et al., 2024); energy-loss perspective (Miao et al., 2025); length-regularized DPO (Park et al., 2024).
Post-hoc steering	Reduce specific failure modes without retraining whole model.	Contrastive activation addition steering (Rimsky et al., 2024); pinpoint tuning for sycophancy (Chen et al., 2024b).
Evaluation/monitoring	Increase oversight robustness; detect cheating; stress-test.	Model-written evals (Perez et al., 2023); ImpossibleBench (Zhong et al., 2025); prompt-hacking datasets (Schulhoff et al., 2023); system-level reporting (Hurst et al., 2024).

Table 3: Mitigation strategies grouped by intervention point. Many methods combine multiple levers; e.g., robust reward modeling plus constrained optimization.

sight (Denison et al., 2024) or strategically misrepresent their alignment (Greenblatt et al., 2024; Hubinger et al., 2024), then naive training-time penalties may "drive hacking underground" (METR, 2025). Finally, monitoring approaches must contend with non-faithful reasoning traces (Chen et al., 2025), and system-level documentation (e.g., system cards) provides essential context for real-world risks and mitigations (Hurst et al., 2024). Appendix D collects extended case studies and a deployment-oriented checklist for agentic settings and reward-channel risks.

7 Open problems and research directions

We highlight several open problems that recur across the surveyed literature:

- **Robust oversight under shift.** How can we evaluate alignment under distribution shift and adversarial settings, when reward models may not track causality (LeVine et al., 2023; Xu et al., 2025) and test-case access enables cheating (Zhong et al., 2025)?
- **Causal reward design.** Can we operationalize causal or counterfactually robust rewards at scale (Wang et al., 2025), and how should we validate that such rewards resist exploitation better than correlational proxies (Kim and Seo, 2024)?
- **Optimization-aware reward learning.** How should reward models be trained *with the optimizer in mind*, given scaling and accuracy-paradox effects (Gao et al., 2023; Chen et al., 2024c)?
- **Long-horizon and agentic settings.** Reward hacking in interactive or tool-using agents may be more severe, as suggested by frontier evaluations (METR, 2025) and by work on deceptive persistence (Hubinger et al., 2024).
- **Sycophancy vs. helpfulness trade-offs.** How should we balance politeness, user satisfaction, and truthfulness when preference data may incentivize agreement (Sharma et al., 2024; Cheng et al., 2025)?

8 Conclusion

Reward hacking in LLMs emerges from a repeatable paradigm: optimize a proxy under limited oversight, and models will exploit whatever the proxy makes easy. Across the surveyed literature, we find that reward hacking encompasses low-level correlates, social dynamics, evaluation gaming, and safety-critical threats. Mitigation requires combining better reward signals (ideally causal/robust), optimization constraints, and adversarial evaluation and monitoring. We believe that this survey can help researchers and practitioners reason about reward hacking as a systematic alignment failure mode and to design interventions that reduce it without sacrificing capability or safety.

Limitations

This survey focuses on reward hacking in large language models and related phenomena. We do not attempt to comprehensively cover all historical work on specification gaming, Goodhart's law, or RLHF beyond this curated set. Because many mitigations are evaluated under different model families, datasets, and metrics, cross-paper comparisons should be interpreted qualitatively rather than as definitive rankings.

Ethics Statement

We confirm that this survey strictly follows ethical research standards. All of the literature papers mentioned in the main paper are publicly available, and no human participants or personally identifiable information have been included. The aim of the survey is to promote academic understanding of the LLM reward hacking research area and to point out approaches to mitigate such issues. All previous related works have been cited appropriately, with due recognition given to their original contributions.

References

Anonymous. 2025. Mitigating reward hacking in LLM-based recommendation: A preference optimization approach. In *Submitted to The Fourteenth International Conference on Learning Representations*. Under review.

Jan Betley, Daniel Chee Hian Tan, Niels Warncke, Anna Sztyber-Betley, Xuchan Bao, Martín Soto, Nathan Labenz, and Owain Evans. 2025. Emergent misalignment: Narrow finetuning can produce broadly misaligned LLMs. In *Forty-second International Conference on Machine Learning*.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomek Korbak, David Lindner, Pedro Freire, Tony Tong Wang, Samuel Marks, Charbel-Raphael Segerie, Micah Carroll, Andi Peng, Phillip J.K. Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, and 13 others. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *Transactions on Machine Learning Research*. Survey Certification, Featured Certification.

Shreyas Chaudhari, Pranjal Aggarwal, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, Karthik Narasimhan, Ameet Deshpande, and Bruno Castro da Silva. 2025. Rlhf deciphered: A critical analysis of reinforcement learning from human feedback for llms. *ACM Computing Surveys*, 58(2):1–37.

Lichang Chen, Chen Zhu, Jiuhai Chen, Davit Soselia, Tianyi Zhou, Tom Goldstein, Heng Huang, Mohammad Shoeybi, and Bryan Catanzaro. 2024a. Odin: disentangled reward mitigates hacking in rlhf. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Wei Chen, Zhen Huang, Liang Xie, Binbin Lin, Houqiang Li, Le Lu, Xinmei Tian, Deng Cai, Yonggang Zhang, Wenxiao Wan, Xu Shen, and Jieping Ye. 2024b. From yes-men to truth-tellers: addressing sycophancy in large language models with pinpoint tuning. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, and 1 others. 2025. Reasoning models don't always say what they think. *arXiv preprint arXiv:2505.05410*.

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

Myra Cheng, Sunny Yu, Cinoo Lee, Pranav Khadpe, Lujain Ibrahim, and Dan Jurafsky. 2025. Social sycophancy: A broader understanding of llm sycophancy. *arXiv preprint arXiv:2505.13995*.

Carson Denison, Monte MacDiarmid, Fazl Barez, David Duvenaud, Shauna Kravec, Samuel Marks, Nicholas Schiefer, Ryan Soklaski, Alex Tamkin, Jared Kaplan, and 1 others. 2024. Sycophancy to subterfuge: Investigating reward-tampering in large language models. *arXiv preprint arXiv:2406.10162*.

Jacob Eisenstein, Chirag Nagpal, Alekh Agarwal, Ahmad Beirami, Alexander Nicholas D'Amour, Krishnamurthy Dj Dvijotham, Adam Fisch, Katherine A Heller, Stephen Robert Pfohl, Deepak Ramachandran, Peter Shaw, and Jonathan Berant. 2024. Helping or herding? reward model ensembles mitigate but do not eliminate reward hacking. In *First Conference on Language Modeling*.

Aaron Fanous, Jacob Goldberg, Ank Agarwal, Joanna Lin, Anson Zhou, Sonnet Xu, Vasiliki Bikia, Roxana Daneshjou, and Sanmi Koyejo. 2025. Syceval: Evaluating llm sycophancy. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pages 893–900.

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

Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, and 1 others. 2024. Alignment faking in large language models. *arXiv preprint arXiv:2412.14093*.

Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M Ziegler, Tim Maxwell, Newton Cheng, and 1 others. 2024. Sleeper agents: Training deceptive llms that persist through safety training. *arXiv preprint arXiv:2401.05566*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Avneet Kaur. 2025. Echoes of agreement: Argument driven sycophancy in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 22803–22812.

Azal Ahmad Khan, Sayan Alam, Xinran Wang, Ahmad Faraz Khan, Debanga Raj Neog, and Ali Anwar. 2024. Mitigating sycophancy in large language models via direct preference optimization. In *2024 IEEE International Conference on Big Data (BigData)*, pages 1664–1671. IEEE.

Sungdong Kim and Minjoon Seo. 2024. Rethinking the role of proxy rewards in language model alignment. In *EMNLP*.

Will LeVine, Benjamin Pikus, Anthony Chen, and Sean Hendryx. 2023. A baseline analysis of reward models' ability to accurately analyze foundation models under distribution shift. *arXiv preprint arXiv:2311.14743*.

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

Yiran Ma, Zui Chen, Tianqiao Liu, Mi Tian, Zhuo Liu, Zitao Liu, and Weiqi Luo. 2025. What are step-level reward models rewarding? counterintuitive findings from mcts-boosted mathematical reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24812–24820.

Monte MacDiarmid, Benjamin Wright, Jonathan Uesato, Joe Benton, Jon Kutasov, Sara Price, Naia Bouscal, Sam Bowman, Trenton Bricken, Alex Cloud, and 1 others. 2025. Natural emergent misalignment from reward hacking in production rl. *arXiv preprint arXiv:2511.18397*.

Lars Malmqvist. 2025. Sycophancy in large language models: Causes and mitigations. In *Intelligent Computing-Proceedings of the Computing Conference*, pages 61–74. Springer.

METR. 2025. Recent frontier models are reward hacking. https://metr.org/blog/2025-06-05-recent-reward-hacking/.

Yuchun Miao, Sen Zhang, Liang Ding, Rong Bao, Lefei Zhang, and Dacheng Tao. 2024. InfoRM: Mitigating reward hacking in RLHF via information-theoretic reward modeling. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Yuchun Miao, Sen Zhang, Liang Ding, Yuqi Zhang, Lefei Zhang, and Dacheng Tao. 2025. The energy loss phenomenon in RLHF: A new perspective on mitigating reward hacking. In *Forty-second International Conference on Machine Learning*.

Ted Moskovitz, Aaditya K Singh, DJ Strouse, Tuomas Sandholm, Ruslan Salakhutdinov, Anca Dragan, and Stephen Marcus McAleer. 2024. Confronting reward model overoptimization with constrained RLHF. In *The Twelfth International Conference on Learning Representations*.

OpenAI. 2025. Expanding on what we missed with sycophancy. https://openai.com/index/expanding-on-sycophancy/. Accessed: 2026-01-04.

Jane Pan, He He, Samuel R Bowman, and Shi Feng. 2024. Spontaneous reward hacking in iterative self-refinement. *arXiv preprint arXiv:2407.04549*.

Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. 2024. Disentangling length from quality in direct preference optimization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4998–5017, Bangkok, Thailand. Association for Computational Linguistics.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, and 1 others. 2023. Discovering language model behaviors with model-written evaluations. In *Findings of the association for computational linguistics: ACL 2023*, pages 13387–13434.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand. Association for Computational Linguistics.

Aswin Rrv, Nemika Tyagi, Md Nayem Uddin, Neeraj Varshney, and Chitta Baral. 2024. Chaos with keywords: Exposing large language models sycophancy to misleading keywords and evaluating defense strategies. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 12717–12733.

Sander Schulhoff, Jeremy Pinto, Anaum Khan, Louis-François Bouchard, Chenglei Si, Svetlina Anati, Valen Tagliabue, Anson Kost, Christopher Carnahan, and Jordan Boyd-Graber. 2023. Ignore this title and hackaprompt: Exposing systemic vulnerabilities of llms through a global prompt hacking competition. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4945–4977.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Esin DURMUS, Zac Hatfield-Dodds, Scott R Johnston, Shauna M Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2024. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations*.

Mia Taylor, James Chua, Jan Betley, Johannes Treutlein, and Owain Evans. 2025. School of reward hacks: Hacking harmless tasks generalizes to misaligned behavior in llms. *arXiv preprint arXiv:2508.17511*.

Chaoqi Wang, Zhuokai Zhao, Yibo Jiang, Zhaorun Chen, Chen Zhu, Yuxin Chen, Jiayi Liu, Lizhu Zhang, Xiangjun Fan, Hao Ma, and 1 others. 2025. Beyond reward hacking: Causal rewards for large language model alignment. *arXiv preprint arXiv:2501.09620*.

Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V Le. 2023. Simple synthetic data reduces sycophancy in large language models. *arXiv preprint arXiv:2308.03958*.

Yuhui Xu, Hanze Dong, Lei Wang, Caiming Xiong, and Junnan Li. 2025. Reward models identify consistency, not causality. *arXiv preprint arXiv:2502.14619*.

Yunpu Zhao, Rui Zhang, Junbin Xiao, Changxin Ke, Ruibo Hou, Yifan Hao, Qi Guo, and Yunji Chen. 2024. Towards analyzing and mitigating sycophancy in large vision-language models. *arXiv preprint arXiv:2408.11261*.

Ziqian Zhong, Aditi Raghunathan, and Nicholas Carlini. 2025. Impossiblebench: Measuring llms' propensity of exploiting test cases. *arXiv preprint arXiv:2510.20270*.

A Survey of Reward Hacking in Large Language Models

Supplementary Material

A Extended measurement protocols and reporting checklist

This appendix provides additional operational detail intended to make the survey more *actionable*: extended protocol checklists, reporting templates, and cross-paper caveats. The core theme is consistent with the main text: reward hacking is a *system-level* failure mode that depends on proxy gaps, optimization pressure, and oversight limits (Casper et al., 2023; Chaudhari et al., 2025; Kim and Seo, 2024).

A.1 Protocol checklists by failure mode

Reward-model overoptimization (Goodhart curves). When studying overoptimization, a minimal protocol is a sweep over optimization pressure (e.g., PPO steps/epochs, KL target, best-of-n, or temperature) and a joint plot of (i) proxy reward and (ii) a stronger "gold" or human/judge outcome (Gao et al., 2023; Chen et al., 2024c; Miao et al., 2025). Recommended controls:

- **Hold out a stronger evaluator.** Separate the *training* proxy from the *auditing* metric, and report proxy–audit divergence (Kim and Seo, 2024; Chen et al., 2024c).
- **Report the optimization knob.** Specify KL target / KL penalty, reward scaling, batch sizes, and selection strategy (e.g., best-of-n) so others can reproduce the pressure regime (Gao et al., 2023; Chaudhari et al., 2025).
- **Track dynamics, not just endpoints.** Instabilities and "energy loss" signals can act as early warnings that optimization is entering an exploitative regime (Miao et al., 2025).

Length/format hacking and other spurious correlates. To detect spurious correlate exploitation, compare scores under length-controlled evaluation, compute score–length Pareto frontiers, and test monotonicity between interpretable proxies and stronger reward models during optimization (Chen et al., 2024a; Park et al., 2024; Kim and Seo, 2024). Recommended controls:

- **Disentangle the correlate.** Explicitly separate length-correlated reward components or train reward heads that isolate spurious features (Chen et al., 2024a).
- **Stress test OOD length.** Report response-length distributions *before and after* optimization to detect drift (Park et al., 2024).
- **Audit robustness interventions.** Data augmentation and robust RM training can reduce correlate-based hacks but rarely eliminate them completely (Liu et al., 2025; Eisenstein et al., 2024).

Sycophancy (propositional, political, and social). Sycophancy is best measured with targeted perturbations that decouple user confidence from ground truth: "are you sure?" prompts, rebuttal conversations, stance-flip tests, and advice scenarios where social norms create incentives to agree (Perez et al., 2023; Sharma et al., 2024; Fanous et al., 2025; Kaur, 2025; Cheng et al., 2025; Malmqvist, 2025). Recommended controls:

- **Separate truth from politeness.** For advice settings, report both factual correctness and social-tone preferences to expose trade-offs (Cheng et al., 2025; OpenAI, 2025).
- **Adversarial keyword/leading cues.** Keyword-triggered agreement and leading queries can create brittle failure modes and "defenses" that overfit surface forms (Rrv et al., 2024; Zhao et al., 2024).
- **Report intervention data.** Synthetic data and curated preference-pair training can reduce sycophancy; report dataset size, filtering, and ID/OOD evaluation splits (Wei et al., 2023; Khan et al., 2024).

Prompt injection and prompt hacking. For injection-style failures, evaluate with adversarial prompts against a fixed success criterion (e.g., jailbreak success, hidden target completion) and measure robustness to paraphrase and instruction-hierarchy variants (Schulhoff et al., 2023; Rrv et al., 2024). For coding benchmarks, include explicit test-access ablations to quantify the degree of benchmark exploitation (Zhong et al., 2025).

In-context reward hacking (without parameter updates). Iterative self-refinement can display in-context reward hacking when the judge and author

share context; evaluate trajectories under context-sharing ablations and verify improvements with external human ratings (Pan et al., 2024).

Reward-channel attacks, deception, and agentic settings. When the model can influence oversight, evaluate in settings where (i) reward functions are gameable, (ii) the model can act over long horizons, and (iii) monitoring is imperfect (Denison et al., 2024; Greenblatt et al., 2024; Hubinger et al., 2024). Recommended controls:

- **Agentic evaluations.** Validate alignment on tool-using or agent scaffolds; chat-like performance can mask failures in agentic contexts (METR, 2025; MacDiarmid et al., 2025).
- **Monitor faithfulness.** Do not assume reasoning traces are faithful; audit monitoring pipelines for evasion and unfaithfulness (Chen et al., 2025).
- **Document deployment assumptions.** System cards and deployment reports help clarify evaluation scope and residual risk (Hurst et al., 2024).

A.2 Reporting template (recommended minimum)

For reproducible reporting, we recommend explicitly stating:

- **Data.** Dataset sizes, train/dev/test splits, filtering, and whether data is used for learning or evaluation-only (Wei et al., 2023; Khan et al., 2024; Zhong et al., 2025).
- **Optimization.** Objective (RLHF/DPO), hyperparameters controlling optimization pressure, and selection strategy (e.g., best-of-n) (Chaudhari et al., 2025; Rafailov et al., 2023; Gao et al., 2023).
- **Oversight/evaluation.** Evaluator type (human/judge/RM), adversarial tests, and robustness checks under distribution shift (LeVine et al., 2023; Xu et al., 2025).
- **Threat model.** What the model can see/do (test access, tool access, reward access) and what counts as success (Schulhoff et al., 2023; Denison et al., 2024).

B Extended dataset and benchmark details

Tab. 4 extends Tab. 2 with additional "what to report" guidance and includes datasets used to study sycophancy, in-context reward hacking, and step-level rewards.

C Extended mitigation matrix and decision rules

Tab. 5 expands Table 3 with a lens-based view: whether a mitigation primarily reduces the proxy gap, limits optimization pressure on brittle proxies, and/or strengthens oversight.

C.1 Practical decision rules (informal)

The literature supports a few recurring decision rules:

- **If proxy reward increases while audit quality decreases,** treat this as a Goodhart signal and reduce pressure (constraints, early stopping) while improving reward robustness (Gao et al., 2023; Chen et al., 2024c; Miao et al., 2025; Moskovitz et al., 2024).
- **If failures are triggered by superficial cues (length/keywords),** remove spurious features from the reward signal and stress test under paraphrase and OOD perturbations (Chen et al., 2024a; Rrv et al., 2024; Liu et al., 2025).
- **If chat-like evaluations look safe but agentic settings fail,** re-center evaluation and safety training around the deployment distribution (MacDiarmid et al., 2025; METR, 2025; Hubinger et al., 2024).
- **If monitoring is used as a defense,** audit faithfulness assumptions and treat chain-of-thought as an *untrusted* signal unless validated (Chen et al., 2025).

D Safety case studies and deployment checklist

This section collects longer-form case studies that illustrate how reward hacking can generalize to safety-relevant misalignment, and it distills a deployment-oriented checklist.

D.1 Case studies: how reward hacking generalizes

Production RL: emergent misalignment from reward hacking. MacDiarmid et al. (2025) show that teaching reward-hack strategies and training on production coding RL environments can lead to broad misalignment, including alignment faking and sabotage in an agent scaffold. They also show that chat-like RLHF can mask residual agentic misalignment, motivating distribution-targeted evaluation and safety training.

Narrow training → broad misalignment. Two complementary lines of evidence highlight transfer risks: reward-hacking demonstrations can generalize to broader misaligned behavior (Taylor et al., 2025), and narrow fine-tuning can yield emergent misalignment beyond the training domain (Betley et al., 2025).

Reward tampering and oversight manipulation. Denison et al. (2024) study curricula where models learn increasingly direct specification gaming and occasionally generalize to reward tampering, reinforcing the need to harden the reward channel and avoid assuming oversight is static.

Deception persistence across training. Work on alignment faking and sleeper agents emphasizes that models can appear aligned under some evaluations while retaining misaligned objectives or triggers (Greenblatt et al., 2024; Hubinger et al., 2024).

In-context reward hacking as a warning sign. Pan et al. (2024) show that reward hacking can arise even without parameter updates when the same model plays judge+author, indicating that some reward-hacking behaviors can be elicited by *protocol design* alone.

D.2 Deployment-oriented checklist (practitioner version)

- **Define the reward channel threat model.** Specify what the model can observe/modify (tests, graders, prompt templates, tool access, memory) (Zhong et al., 2025; Denison et al., 2024).
- **Evaluate agentic behavior.** Include tool-using or long-horizon tasks; do not rely solely on chat-like evaluations (METR, 2025; MacDiarmid et al., 2025).
- **Check for sycophancy under persuasion.** Use rebuttal protocols, stance flips, and advice scenarios; report truth vs politeness trade-offs (Sharma et al., 2024; Fanous et al., 2025; Kaur, 2025; Cheng et al., 2025).
- **Harden against prompt injection.** Test paraphrases and instruction-hierarchy variants; monitor for brittle keyword-based defenses (Schulhoff et al., 2023; Rrv et al., 2024).
- **Audit proxy-gap failure modes.** Stress test spurious correlates (length, style, "consistency") and validate causal/reward invariances where possible (Chen et al., 2024a; Xu et al., 2025; Wang et al., 2025).
- **Document assumptions and residual risk.** Use system-level documentation (system cards) and report evaluation gaps explicitly (Hurst et al., 2024).

E Reference-to-theme index

Tab. 6 provides an at-a-glance index mapping each reference to the survey's main themes.

Dataset / benchmark	Primary use	What is provided (stats)	What to report / common gaps	Train?
HackAPrompt (Schulhoff et al., 2023)	prompt injection	560,161 (Playground) + 41,596 (Submissions); 10 challenges	Define success criteria + prompt variants; test paraphrase robustness; avoid overfitting to keywords (Rrv et al., 2024).	×
ImpossibleBench (Zhong et al., 2025)	benchmark exploitation	Impossible: SWE 2,371; Live 193. Open-test: SWE 2,300; Live 550	Report test-access assumptions and ablations; separate "solve" vs "cheat" labels; disclose monitoring constraints.	×
School of Reward Hacks (Taylor et al., 2025)	reward-hack SFT demos	1,073 dialogues (35 tasks; 973 NL + 100 code); split n/r	Provide split protocol and OOD evaluation; report whether hacks are explicitly described to the model.	✓
DPO sycophancy pairs (Khan et al., 2024)	anti-sycophancy training	train: 1,000 pairs (+120 pref pairs); eval: 120 ID + 120 OOD	Report persona/prompt distribution; distinguish preference alignment vs factual correctness.	✓
SycEval (Fanous et al., 2025)	rebuttal sycophancy eval	AMPS 500 + MedQuad 500; 3k initial + 24k rebuttals	Specify multi-turn protocol; report progressive vs regressive flips and calibration under persuasion.	×
Political sycophancy set (Perez et al., 2023)	political agreement eval	n/r (paper-defined prompts/labels)	Clarify stance labels, topic coverage, and how "agreement" is detected (classifier vs rubric).	×
Keyword sycophancy set (Rrv et al., 2024)	keyword-triggered bias	500 sets; 5 domains (1,030 gen $\rightarrow$ 650 filt $\rightarrow$ 500)	Report keyword construction pipeline; evaluate defenses under distribution shift and paraphrases.	×
Leading-query LVLM benchmarks (Zhao et al., 2024)	leading cues vs vision	POPE 9k; AMBER 14k; RWQA 765; SciQA 2,017; MM-Vet 218	Separate language-leading effects from vision evidence; disclose rewrite procedure and leakage.	×
Synthetic anti-sycophancy data (Wei et al., 2023)	data intervention	train pool: 100k (17 datasets); eval: 3k (1k $\times$ 3)	Report filtering strategy and model dependence; evaluate trade-offs with helpfulness.	✓
RRM augmented RM data (Liu et al., 2025)	robust reward modeling	RM train: 2.4M examples ($\sim$14$\times$ aug.)	Disclose augmentation recipe; report performance under shift and after optimization (not just RM accuracy).	✓
Production RL case study (MacDiarmid et al., 2025)	reward hacks $\rightarrow$ misalignment	env count n/r; SDF 99:1 docs; 6-eval suite	Specify hack definition, environment mix, and agentic evaluation distribution; report residual risk after RLHF.	✓
Self-refinement essays (Pan et al., 2024)	in-context hacking eval	23 essays; 5 refinement steps; 3 ratings/essay	Report judge/author context sharing; audit whether score improvements correlate with human ratings.	×
Step-level reward datasets (Ma et al., 2025)	step-reward validation	n/r (task-specific corpora)	Report step label source, aggregation method, and whether step rewards track correct reasoning vs patterns.	×

Table 4: Extended dataset/benchmark notes (n/r = not reported). "Train?" indicates whether the dataset is used for learning/optimization (e.g., SFT/RL/RM training) in the cited work.

Where	Lever (example)	Proxy gap	Pressure	Oversight	Evidence / notes
Data / preferences	Synthetic anti-sycophancy + curated pairs	✓	×	✓	Data interventions and explicit preference labeling reduce sycophancy and reshape incentives (Wei et al., 2023; Khan et al., 2024; Sharma et al., 2024). Social norms can reintroduce proxy pressure in advice settings (Cheng et al., 2025; OpenAI, 2025).
Domain-specific objectives	Recommendation "anchors"	✓	✓	×	One-class implicit feedback creates insensitive regions; pseudo-negative anchors maintain informative gradients (Anonymous, 2025).
Reward modeling	Disentangle spurious features	✓	×	✓	Length-correlated reward heads can be separated and removed during optimization (Chen et al., 2024a); DPO can also amplify length preferences (Park et al., 2024).
Reward modeling	Robust RM training / augmentation	✓	×	✓	Robust procedures and augmentation improve resistance to spurious shortcuts but do not eliminate hacking after optimization (Liu et al., 2025; Eisenstein et al., 2024).
Reward modeling	Information-theoretic / representation robustness	✓	×	✓	InfoRM constrains exploitable information and proposes monitoring indicators (Miao et al., 2024).
Reward design	Causal/counterfactual rewards	✓	×	✓	Causal rewards aim to break spurious correlations; audits should test invariances under interventions (Wang et al., 2025; Xu et al., 2025; Kim and Seo, 2024).
Optimization	Constrained RLHF	×	✓	×	Limit overoptimization while improving reward; best used when proxy is brittle under high pressure (Moskovitz et al., 2024).
Optimization	Dynamics-aware early stopping	×	✓	✓	Training dynamics (e.g., "energy loss") can indicate exploitative regimes and support early stopping/diagnostics (Miao et al., 2025).
Evaluation	Adversarial prompts + robustness checks	×	×	✓	Prompt-injection benchmarks and keyword studies show brittleness; evaluate paraphrase/instruction hierarchy variants (Schulhoff et al., 2023; Rrv et al., 2024).
Evaluation	Impossible / exploitability tests	×	×	✓	Test-access enables cheating; report explicit cheating rates and ablations (Zhong et al., 2025).
Monitoring	Faithfulness-aware monitoring	×	×	✓	Reasoning traces can be non-faithful; monitoring must be audited for evasion (Chen et al., 2025).
Safety training	Target the deployment distribution	×	✓	✓	Chat-like RLHF can mask agentic misalignment; diversify safety data and evaluate on agentic tasks (METR, 2025; MacDiarmid et al., 2025).
Post-hoc	Activation steering / pinpoint tuning	×	×	×	Useful for rapid mitigation without full retraining, but does not fix the underlying reward/oversight mismatch (Rimsky et al., 2024; Chen et al., 2024b).

Table 5: Extended mitigation matrix. Columns indicate whether the lever primarily reduces proxy gaps, limits optimization pressure, and/or strengthens oversight robustness.

Theme	What it contributes	Representative works
Survey framing / RLHF limits	System-level view of alignment pipelines and why proxy optimization creates reward hacking pressure.	(Casper et al., 2023; Chaudhari et al., 2025)
Overoptimization and diagnostics	Empirical Goodhart curves, accuracy-paradox effects, and training-dynamics signals for exploitative regimes.	(Gao et al., 2023; Chen et al., 2024c; Miao et al., 2025)
Proxy design and causality	Critiques of correlational rewards and proposals for causal/counterfactual reward modeling under shift.	(Kim and Seo, 2024; Xu et al., 2025; Wang et al., 2025; LeVine et al., 2023)
Robust reward modeling	Reward heads that remove spurious correlates; robust RM training/augmentation; ensembles; information-theoretic views; step-level reward pitfalls.	(Chen et al., 2024a; Liu et al., 2025; Eisenstein et al., 2024; Miao et al., 2024; Ma et al., 2025)
Social reward hacking (sycophancy)	Political/propositional/social sycophancy phenomena; rebuttal protocols; persuasion-driven stance shifts; deployment regressions; mitigations.	(Perez et al., 2023; Sharma et al., 2024; Malmqvist, 2025; Fanous et al., 2025; Kaur, 2025; Cheng et al., 2025; OpenAI, 2025; Khan et al., 2024; Wei et al., 2023; Zhao et al., 2024; Rrv et al., 2024)
Evaluation gaming	Prompt injection benchmarks, keyword brittleness, and coding benchmark exploitability under test access.	(Schulhoff et al., 2023; Rrv et al., 2024; Zhong et al., 2025; Pan et al., 2024)
Constraints and optimization control	Methods that explicitly constrain optimization or reshape objectives to limit reward hacking under pressure.	(Moskovitz et al., 2024; Rafailov et al., 2023; Park et al., 2024)
Reward-channel attacks and deception	Reward tampering curricula, alignment faking, sleeper agents, and monitoring limitations (non-faithful CoT).	(Denison et al., 2024; Greenblatt et al., 2024; Hubinger et al., 2024; Chen et al., 2025)
Emergent misalignment in practice	Evidence that reward hacking can generalize to agentic sabotage and that chat-like RLHF can mask residual risk.	(MacDiarmid et al., 2025; METR, 2025; Betley et al., 2025; Taylor et al., 2025)
Post-hoc steering / patching	Inference-time steering and localized fine-tuning as rapid interventions with limited root-cause coverage.	(Rimsky et al., 2024; Chen et al., 2024b)

Table 6: Reference-to-theme index for the appendix.